

12-2011

The Valuation of User-Generated Content: A Structural, Stylistic and Semantic Analysis of Online Reviews

Noi Sian KOH

Singapore Management University, noisian.koh.2006@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/etd_coll

Part of the [Databases and Information Systems Commons](#), and the [E-Commerce Commons](#)

Citation

KOH, Noi Sian. The Valuation of User-Generated Content: A Structural, Stylistic and Semantic Analysis of Online Reviews. (2011). 1-168. Dissertations and Theses Collection (Open Access).

Available at: https://ink.library.smu.edu.sg/etd_coll/78

This PhD Dissertation is brought to you for free and open access by the Dissertations and Theses at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Dissertations and Theses Collection (Open Access) by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

**The Valuation of User-Generated Content:
A Structural, Stylistic and Semantic Analysis of Online Reviews**

by

KOH Noi Sian

Submitted to School of Information Systems in partial fulfillment
of the requirements for the Degree of Doctor of Philosophy in Information Systems

Dissertation Committee

Nan HU (Committee Chair)
Assistant Professor of Information Systems
Singapore Management University

Jing JIANG
Assistant Professor of Information Systems
Singapore Management University

Steven MILLER
Professor of Information Systems (Practice)
Singapore Management University

Srinivas K REDDY
Professor of Marketing
Singapore Management University

Singapore Management University
2011

Copyright (2011) KOH Noi Sian

Abstract

The ability and ease for users to create and publish content has provided vast amount of online product reviews. However, the amount of data is overwhelmingly large and unstructured, making information difficult to quantify. This creates challenge in understanding how online reviews affect consumers' purchase decisions. In my dissertation, I explore the structural, stylistic and semantic content of online reviews. Firstly, I present a measurement that quantifies sentiments with respect to a multi-point scale and conduct a systematic study on the impact of online reviews on product sales. Using the sentiment metrics generated, I estimate the weight that customers place on each segment of the review and examine how these segments affect the sales for a given product. The results empirically verified that sentiments influence sales, of which ratings alone do not capture. Secondly, I propose a method to detect online review manipulation using writing style analysis and assess how consumers respond to such manipulation. Finally, I find that societal norms have influence on posting behavior and significant differences do exist across cultures. Users should therefore exercise care in interpreting the information from online reviews. This dissertation advances our understanding on the consumer decision making process and shed insight on the relevance of online review ratings and sentiments over a sequential decision making process. Having tapped into the abundant supply of online review data, the results in this work are based on large-scale datasets which extend beyond the scale of traditional word-of-mouth research.

Table of Contents

Acknowledgements	iii
Chapter 1. An Introduction of Online Reviews	1
1.1. Background and Motivation	1
1.1.1. Design: Structure and characteristics of online reviews	3
1.1.2. Value: Underlying mechanism and economic value	8
1.1.3. Behavior: Posters' motives and behavior	9
1.2. Objectives.....	10
1.2.1. Interplay: Differing forces between sentiments and ratings	10
1.2.2. Manipulation: Posters' motives and effect on sales.....	12
1.2.3. Cultural forces: Reporting behavioral patterns across cultures	12
1.3. Dissertation Questions and Contributions	12
1.4. Implications for Research and Practice	15
1.4.1. Managerial Perspective.....	15
1.4.2. Readers' Perspective	17
1.5. Organization of the Thesis	18
Chapter 2. Background and Related Work.....	19
2.1. Word of Mouth.....	20
2.1.1. Quantitative Analysis – A Marketing Perspective.....	20
2.2. Mining Online Reviews	23
2.2.1. Qualitative Analysis – A Sentiment Mining Perspective	23
2.2.2. Scale and the need for automation.....	25
2.3. Gaps in Current Research for Online Reviews.....	28
2.3.1. Structure and Characteristics	28
2.3.2. Assumptions and Bias	29
Chapter 3. Methodology and Measurements	34
3.1. Technology & systems for scalable, semi-automated execution.....	35
3.1.1. System designs for sentiment extraction and mining	36
3.1.2. Sentiment measurements	44
3.2. Other measures on the stylistic content.....	47
3.2.1. Stylistic measurements for manipulation detection	47
3.3. Dataset Description	48
3.3.1. Amazon cross-sectional dataset.....	49
Chapter 4. Impact of Sentiments on Sales.....	53

4.1.	Conceptual framework	54
4.2.	Online Reviews Format: Impact of Sentiments on Sales	58
4.3.	Robustness Check	66
4.3.1.	On a different sentiment mining technique.....	66
4.3.2.	On a different dataset.....	69
4.4.	Implications.....	71
Chapter 5.	Manipulation of Online Reviews	73
5.1.	Manipulation detection.....	74
5.1.1.	Existence of manipulation activity	75
5.1.2.	Randomness of writing style	79
5.1.3.	Wald-Wolfowitz (Runs) test.....	82
5.2.	Robustness Check	83
5.2.1.	Evidence of manipulation discovered by Runs test.....	83
5.3.	Impact of Manipulation on Sales.....	86
5.4.	Implications.....	93
Chapter 6.	Under-reporting Bias and Online Reviewers' Behavior.....	95
6.1.	Conceptual Development	101
6.1.1.	Attitude.....	102
6.1.2.	Social Norms	102
6.1.3.	Motivation	106
6.2.	Research Methodology.....	107
6.2.1.	Cross-Cultural Data	107
6.2.2.	Experimental Calibration with Survey Data.....	110
6.2.3.	Graphical Data Analysis.....	111
6.3.	Impact of Attitude and Social Norms on Rating Behavior.....	117
6.4.	Motivation for Writing Online Reviews.....	120
6.5.	Under-reporting Bias	123
6.6.	Robustness Check	128
6.6.1.	On a different dataset.....	128
6.7.	Implications.....	130
Chapter 7.	Discussion and Conclusion.....	134
7.1.	Future Work and Extensions	136
REFERENCES.....		145
Appendix.....		158

Acknowledgements

This dissertation would not have been possible without the guidance, support and help from many people. I wish to express my heartfelt gratitude to my advisor Asst. Professor Hu Nan, my committee members Asst. Professor Jiang Jing, Professor Steven Miller and Professor Srinivas Reddy for your patient guidance and supervision in the course of this dissertation. Your constructive advice has guided me on the right track and made me sustain through this laborious research, without which this dissertation might never have been completed. To Jason, Narayan, Youngsoo, Prof Kam and Chen Bin, thank you for your suggestions and help in one way or another. Many thanks to Professor Eric Clemons for spending a generous amount of time in teaching and guiding me.

To all my teachers who have taught me in my coursework and along my years of PhD studies. You have provided tremendous inspiration and mental support in the early stage of my PhD studies. I am so glad I came to SIS and took your courses. Hopefully, I can be as inspiring to my students in future.

To my PhD classmates Han Jin, Yan Qiang, Fu Na, Meiqun, Jianhui and many whom I have met in SMU. Thank you for your immeasurable support, I cannot imagine going through my PhD without you. Last but not least, my appreciation to my family for being there when I needed them.

Although the materialization of this dissertation is only possible with all the help from those cited above, the responsibilities for any errors and omissions in this work are solely mine.

Chapter 1.

An Introduction of Online Reviews

1.1. Background and Motivation

Social media has radically changed the way we communicate and share information on the web. The shift from a one way communication to a conversation style interaction has led to the generation of online reviews, pictures, videos and audio. The content produced in social media is often referred to as “user-generated content”. As opposed to professionally edited text (news sites and magazine articles for instance), user-generated content contributes to a rapid growth of content present on the Web today. Earlier, when an individual is thinking of which book to buy, there were very few sources of information to help the consumer make his or her purchase decision. Now, one can simply go to online review sites to gather information on prior customers’ sentiments of the product and then make their purchase decision based on the online reviews read.

Since there are large amount of customer reviews¹ available, they serve as an informative indicator of customers' sentiments and satisfaction. Bickart and Schindler (2001), drawing upon intuition from the rich literature on persuasion, hypothesize that Internet forum content may be more persuasive than other traditional sources of information (such as marketer-generated content) since the reported experiences of peer consumers have the ability to generate empathy among readers and may appear more credible, trustworthy, and relevant. And it is this capacity of persuasion - one of the defining features of online reviews that provide the strongest possible reason for studying the value of online reviews. Nielsen in a large scale (26,000 participants) global study in April 2007 found that 78% of participants trust recommendations from other consumers². Power Reviews in a November 2007 survey found that 68% of the online shoppers read at least four product reviews before purchasing³.

Evidently, user-generated online reviews have become an important source of information to consumers in their search, evaluation and choice of products. A comprehensive understanding of online reviews and the sentiments⁴ expressed in them is of high importance, because these online reviews can be a very good indicator of the product's future sales performance. Thus, the motivating question that has guided this thesis is: "What is the value of online reviews?" The word 'value' here may refer to estimating the impact of online reviews on product sales (for firms) or it

¹ In this dissertation, references will be made to "online reviews", "online/electronic word-of-mouth", "customer reviews", "consumer reviews" and "online customer/consumer review". These terms will be used interchangeably throughout this thesis.

² http://www.nielsen.com/media/2007/pr_071001.html

³ http://www.powerreviews.com/social-shopping/news/press_breed_11122007.html

⁴ Sentiments refer to the positive or negative opinions expressed in customer reviews. In marketing literature, it is called valence. The measurement of valence in marketing, however, is either based on numeric rating or human coded.

may refer to the informative worth of online reviews which help users of online reviews to gauge the true perceived quality of products.

This dissertation is based on two key observations:

1. Understanding the value of online reviews requires identifying the structure, properties, assumptions and underlying bias in online reviews.
2. In terms of the online review content, ratings and sentiments may play different role in the decision-making process. As the decision-making process may consist of various stages such search, evaluation and purchase, the main focus for this dissertation however, will be at the purchase stage. (The understanding of how online reviews affect search and evaluation will be interesting future work to be explored, see Appendix).

The next subsection covers some current misconception and background which will motivate the issues to be discussed in Section 1.1.2 and Section 1.1.3. Then, Section 1.2 presents the objectives of the dissertation and Section 1.3 lists out the specific research questions and contributions of the 3 main studies in this dissertation. Section 1.4 presents the implications of the findings and finally Section 1.5 outlines the organization of the dissertation.

1.1.1. Design: Structure and characteristics of online reviews


Misconception 1: Rating provides a summary of the text sentiments.

Although numerous research use ratings as representative summaries of the text (e.g. Godes and Silva 2009; Moe and Trusov 2009; Duan, Gu and Whinston 2008; Dellarocas, Zhang and Awad 2008; Li and Hitt 2008; Chevalier and Mayzlin 2006;

Clemons, Gao and Hitt 2006; Hu, Pavlou and Zhang 2006; Godes and Mayzlin 2004), ratings may not be good summaries of consumers' real sentiments. Often, we can assume that ratings are a numeric quantification of the text and their valences are consistent, this may not always be valid. Figure 1.1.1a shows some instances in which ratings are not supported by text. Although both reviews are given 3-stars rating out of 5-stars, the sentiments expressed in the first two sentences as clearly illustrates a positive or even high sentiment scores.

Customer Review

14 of 15 people found the following review helpful:

 **Recommended Reading**, April 21, 2000

By **A Customer**

This review is from: **The Love of the Last Tycoon (Paperback)**

Don't be misled by the three-star rating. This was clearly going to be a four- or five-star book, except that Fitzgerald died after completing only the first 17 of 30 intended episodes. The writing is his most economical since *Gatsby*, and the setting of Hollywood provides good fodder for Fitzgerald's recurring theme of scandal among the wealthy or celebrated. The story is related, for the most part, by a woman, the daughter of a well-known producer, about events that occurred five years earlier, when she was in college and in love with a dynamic young producer named Monroe Stahr. Though she loves him from a distance, her somewhat obsessive interest in the man is a useful way to relate his story. The writing was at times vintage Fitzgerald, sometimes recognizably unfinished, but always worth the experience. The notes, letters and outlines included in the version I read were extremely interesting and worth their inclusion. This is a book that I don't think anyone can read without saying, "I wish he had finished this." This is also a book that I recommend to anyone who appreciates and enjoys the writing of F. Scott Fitzgerald.



Lori rated it 

Aug 24, 2009

I'm having such an easier time getting into this than *Here Be Dragons*.

Don't be misled by the 3 star rating, this was a very good book! I notice another review says "I'm glad I read it, and I'm glad it's over", which is how I feel.

The history was fascinating, I knew nothing about Maude and the civil war in 12 century England. Both she and Stephen, who usurped her crown, were extremely well depicted and fully fleshed out into real people. And when one thinks of a usurper, one usually thinks of a power hungry egomaniac, yet Stephen was far from that - he was so likable. A sweet nice man, a baaaaad king really because he was so nice. Likewise Maude, so admirable, had her faults as well, exacerbated by being a strong intelligent independent woman in a world where women are supposed to be just the opposite - well, which is why her queenship was snatched from her, a WOMAN as the consecrated monarch?

I also loved the way Mathilda, Stephen's wife, grew into herself. There were many characters in this history, all of whom were made very real and complex.

Figure 1.1.1a: Examples of rating-sentiment inconsistency.

Hence, to fully understand the influence of online reviews, we first have to understand its structure/design and find out the relative importance of each segment, yet this is an issue which had been overlooked in current studies. Figure 1.1.1b shows an example of an online customer review in Amazon.com. For each review, there is a review date, reviewer's name /nickname, number of people who found the review helpful, the numeric rating and text review provided by a reviewer. Specifically in the text portion of an online review, there are two sub-components 1) the review title which is a highlighted short summary of the customers' overall sentiments and 2) the review content / body which contains the detailed comments of the customers' evaluation of the product. Such a design is also found in other websites such as IMDB.com (see Figure 1.1.1c). The bolded font of the title suggests that sentiment expressed in this portion of the review summarizes the overall level of satisfaction experienced by the customer. Since it is more eye-catching than other textual content, it may also have a higher influence on sales than the sentiments expressed in the review body. Also most of studies have either investigated either the impact of text or ratings, but not their interplay. Therefore, in consideration of this design, we are interested in conducting a systematic study on how each segment of an online review and the interaction between rating and sentiments will affect sales.

7 of 7 people found the following review helpful:

★★★★★ Interesting, Fun, Deeply Thought-Provoking, June 22, 2008

By [Matt Humphrey](#) (San Francisco, CA USA) - [See all my reviews](#)

This review is from: [Sway: The Irresistible Pull of Irrational Behavior \(Hardcover\)](#)

The Brammans do an excellent job showcasing the irrational behavior all around us. Whether you're a doctor, venture capitalist, teacher, or even a college football coach, there are subtle psychological cues driving you to engage in irrational behaviors that can have a significant negative impact on your life. Reading the anecdotes, one might wonder 'how can anyone ever do that?' The book's close inspection of many different situations shows us that we all do it, and in fact, most of us are guilty of irrationality every single day. 'Sway' lifts the mystery behind these subtleties of irrational thinking and allows us to be more critical of ourselves so we can understand really what is driving the decisions we make day in and day out.

Overall, 'Sway' is a great read. It's very well-written, fast-moving, inherently entertaining, insightful, and just downright fun. It will leave you in a healthy state of self-reflection and critical thinking of the world around you.

Figure 1.1.1b: Screenshot of Customer Reviews in Amazon.com

26 out of 44 people found the following review useful:

Visually stunning, subtle, dreamy, 5 February 2010

★★★★★★★★

Author: [chadelle \(chadelle_jb@aliceadsl.fr\)](#) from france

Absolutely stunning. Simply the most beautiful underwater imagery I've ever seen. It's hard to remain not too affected when talking about ecology. Here, the off screen speech is quite subtle, not too naive and not boring, because sparingly used, which leaves long lapse of dreamy sequences, without a word. Technically, it's easily one of the best documentary ever made. The camera work and photography are incredible, the montage is very effective, alternating slow and fast paced sequences. The score is not too obtrusive. There is a very striking scene, which reminds me the nautical funerals of Laetitia in "Les Aventuriers" by Robert Enrico, if you see what I'm referring to, you will easily notice it, and I assure this scene will stick to your mind for days... Visually stunning, subtle, very recommended.

Figure 1.1.1c: Screenshot of Customer Reviews in IMDB.com

Misconception 2: Online reviews are written by customers

Generally, it is assumed that online reviews are posted by customers who have bought and tried the product. However, in many online review sites, it is optional for reviewer to disclose their real name. Since participants of online review communities can assume any identities or choose to be anonymous, marketers are able to disguise their promotion as consumer recommendations. For example, due to software errors, Amazon.com's Canadian site accidentally revealed the true identities of some of its book reviewers. And it was found that a sizable proportion of these reviews were actually written by the book's own publishers, authors and their friends or relatives (Harmon 2004). The music industry for instance is known to hire professional

marketers who surf various online chat rooms and fan sites to post positive comments on new albums (Mayzlin 2006; White 1999). Hence the extent to which online reviews are manipulated has been a question of interest to analysts, researchers and consumers. While there are some recent commitment to investigate online review manipulations through analytical works (Dellarocas 2003 & 2006; Mayzlin 2006), there has been little empirical work in investigating and detecting online review manipulation.

Misconception 3: Online reviews reported match average perceived assessments.

Although user-generated online reviews are a major source of information for consumers to infer product quality, prior research (Hu et al. 2006) has found evidence that online reviews may not be representative of the average perceived assessments due to under-reporting bias. Under-reporting bias is a form of self-selection bias described in the literature on satisfaction (Anderson 1998). Consumers who are very satisfied or very dissatisfied will be more motivated to voice their opinions through reviews and thus are more likely actually to do so. It has been found that under-reporting bias does exist in certain U.S. online review websites (Hu et al. 2006). Hence, the average of reported ratings (created by a small population of those sufficiently motivated to post their reviews) do not match the average of perceived assessments of the general population. Since consumers are becoming increasingly dependent on online reviews to make purchase decisions, it is necessary to find whether under-reporting bias exists across cultures, and whether online consumer rating behavior will yield biased or unbiased estimators of a product's quality in various markets.

Given the background on current misconceptions, the next two subsections explore the issues which will be studied in this dissertation.

1.1.2. Value: Underlying mechanism and economic value

A good number of researchers have focused on examining the value of online consumer reviews on product sales concentrating on numeric ratings that accompany the reviews (e.g. Godes and Mayzlin 2004; Dellarocas et al. 2004; Li and Hitt 2008; Clemons, Gao and Hitt 2006; Dellarocas and Narayan 2006; Hu, Pavlou and Zhang 2006; Chevalier and Mayzlin 2006). However, little work has been done in examining the role on the qualitative aspects of the review (text sentiments) in affecting product sales.

Taking into consideration of Misconception 1, we are interested in conducting a systematic study on how each segment of an online review and the interaction between rating and sentiments will affect sales. While numeric ratings can be viewed as codified assessments on a standardized scale, sentiments expressed in the text provide more tacit, context-specific explanations of the reviewer's feelings, experiences and emotions about the product or service. They could be framed as highly positive, neutral, or negative statements with varying degrees of emotion. Such sentiments provide rich information to their readers and are likely to provide them with a tacit feel, beyond the numeric ratings.

Our stance is that consumer sentiments influence products' sales, which the numeric ratings alone do not capture. It is vital to capture this rich semantic aspect of online reviews to better understand the purchase behavior of online consumer. For these

reasons, we aim to understand how consumers derive quality information from online reviews to make purchase decision.

1.1.3. Behavior: Posters' motives and behavior

Misconception 2 and 3 has presented issues on online review poster's fraudulent motives and under-reporting bias behavior which have not been thoroughly dealt with in current online review literature.

There is growing evidence that consumers are influenced by online reviews before making their purchase decisions (e.g. Senecal and Nantel 2004; Chevalier and Mayzlin 2006), hence it is lucrative for firms to manipulate consumer perceptions by posting anonymous messages that praise their products. And as more firms realize the persuasive power of online reviews, it is expected that more will engage in manipulation practices. Therefore, it is important and timely to understand the impact of such fraud posters' behavior.

Secondly, as each posted online review is an assessment of an individual's perceived quality of a product; such reported quality could be influenced by cultural factors. The behavior of individuals in online networks can be very different, may vary in systematic ways across cultures, and may differ from offline behavior as well. Then, the information in online networks needs to be interpreted carefully before these reviews can be of use to either the community or marketers.

1.2. Objectives

Given the background and motivation presented in Section 1.1, this dissertation contains 3 studies relating to the value of online reviews. This section specifically addresses the objectives for each study.

Given the backdrop, many interesting questions arise on how we can analyze online reviews and study its utility? The first study mainly examines: How do users use the information in online reviews to make their purchase decisions? While the second study presents an online review manipulation detection technique and examines the impact on sales rank if reviews are manipulated. Finally, the third study examines whether the reported average evaluation of the reviews matches the perceive assessment of the population of both raters and non-raters.

1.2.1. Interplay: Differing forces between sentiments and ratings

As with offline search behavior, a consumer searching for product information in cyberspace does not read every relevant online review recommendation before making a purchase decision. Doing so would be nearly impossible given the number of Web sites dedicated to providing consumer reviews and the time pressure consumers often face in searching for and purchasing products. Even a single Web site may contain far more reviews than what consumers can process. How then do consumers select the reviews they read?

Our sense is that customer typically screens through a list of search results and identify a subset to evaluate in greater depth. Based on the search results returned, they become aware of a subset of books which they are interested to click on and evaluated further. Hence, (apart from price and other attributes of the product) we

envisage that average ratings are used in the filtering of search results at the initial stage while sentiments are used to further evaluate the products. Figure 1.2.1 presents a conceptual framework on the relevance of rating and sentiments over the stages of decision making process.

To date, the relevance of rating and sentiments over the stages of decision making process is still unclear, particularly at the purchase stage. Very little work has investigated the interplay of ratings and sentiments (Tsang and Prendergast 2009). Therefore, it is crucial to comprehensively study how consumers derive quality information from online reviews, by taking into consideration the interplay between ratings and sentiment. To achieve this, the *first objective* is to 1) quantify text sentiments; then 2) analyze how ratings, sentiments and their interplay affects consumer's purchase decision.

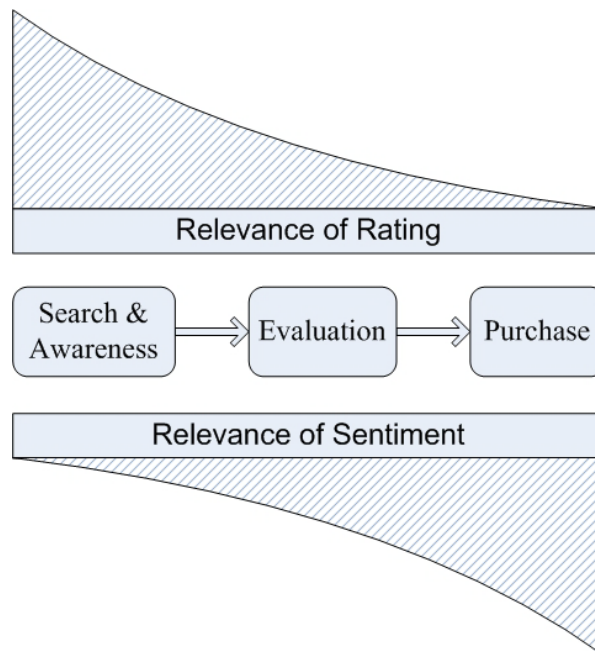


Figure 1.2.1: Stages in the Decision Making Process

1.2.2. Manipulation: Posters' motives and effect on sales

As consumers become increasingly reliant on online reviews to make purchase decision, the sales of the product becomes dependant on the word of mouth that it generates. As a result, there can be attempts by firms to manipulate online reviews of products to increase their sales. Despite the existence of such manipulation, the amount of such activity is unknown, and deciding which reviews to believe in is largely based on the reader's discretion and intuition. Thus, the *second objective* is then to develop a method to detect manipulation of reviews and examine how the manipulation of reviews affects product sales.

1.2.3. Cultural forces: Reporting behavioral patterns across cultures

Under-reporting bias is a form of self-selection bias described in the literature on satisfaction (Anderson 1998) and prior research (Hu et al. 2006) has found evidence that online reviews may not be representative of the general consensus opinions as the posters of online reviews are mostly those who are either very satisfied or very dissatisfied with the product. Since consumers are becoming increasingly dependent on online reviews to make purchase decisions, the *third objective* is to study raters' behaviors to find out whether under-reporting bias exists across cultures, and whether online consumer rating behavior will yield biased or unbiased estimators of a product's quality in various markets.

1.3. Dissertation Questions and Contributions

This dissertation specifically addresses the following research questions for each study:

First study

1. *What is the influence of rating, sentiments and rating-sentiment interplay on sales?*
2. *What is the impact of each segment of an online review on sales?*
3. *Is there a differential role between ratings and sentiments on sales?*

Second study

4. *What is the extent of manipulation amongst online reviews?*
5. *How can such manipulation be detected? What are some of the characteristics that can be used to distinguish between manipulated and non-manipulated reviews?*
6. *What is the impact of manipulation activity on the sales of products?*

Third study

7. *What factors motivate consumers to write online reviews?*
8. *How does culture influence raters' behavior when writing reviews and how cultural differences manifest in differences among ratings?*

My dissertation fills in the gap of current online word-of-mouth research that will support the efforts and advance the understanding of a growing community of scholars in this area. The key contributions are:

First, we conduct a systematic study on the value of online reviews on product sales. Using the sentiment metrics generated, we estimate the weight that customers place on each segment of the review and examine how these segments affect the sales for a given product. The results empirically verified that sentiments influence sales rank, of which ratings alone do not capture. Taking into consideration of the online

review design, the results showed that there is a differential effect of sentiments expressed in the title of the review and the content of the review.

Second, prior work either examines ratings or sentiments and not their differential role. Our experiential survey shed insight on the *relevance* of rating and sentiments over different stages of the consumer decision making process. We find that customers tend to use ratings in their search for information and in filtering of the search results and sentiments to evaluate and make their final choice. The results in this study suggests the relevance of ratings and sentiments may be different over the course of search, evaluation and purchase and we hope that these findings would be an impetus for future research on this timely and important topic.

Third, historical work in this area such as that of Chevalier and Mayzlin (2006), Archak et al (2007), Forman et al. (2008) have used sales rank as the dependent variable in their analyses. We went beyond our existing dataset and obtained the actual point-of-sales data provided by Neilson Bookscan to check the veracity of our results. Our results still hold and the conclusion remains the same i.e. the variance explained by sentiments is much higher than that of ratings; the sentiments in the content impact sales more compared to the sentiments in the title alone; the interaction between ratings and sentiments is statistically significant on sales. This has led us to robust findings that shed light on the effect of online reviews.

Fourth, a statistical method is proposed to detect online review manipulation, and an assessment is made on how consumers respond to such manipulation. In particular, the writing style of reviewers is examined and the effectiveness of manipulating through ratings, sentiments and readability is investigated. Our analysis examines

textual information available in online reviews by combining sentiment mining techniques with readability assessments. We discover that around 10.3% of the products are subjected to online reviews manipulation. In spite of the deliberate use of sentiments and ratings in manipulated products, consumers are only able to detect manipulation through ratings, but not through sentiments.

Finally, we find that societal norms have influence on posting behavior and significant differences do exist across cultures and firms should exercise care in interpreting the information from online reviews.

Having tapped into the abundant supply of online review data, the results in this work are based on large-scale datasets which extends beyond the scale of traditional word-of-mouth research. The rigor of the analyses in each study presents robust and consistent results which are compelling and directional to managers, marketers and users of online reviews.

1.4. Implications for Research and Practice

The empirical analyses in this dissertation have generated useful insights which are supported by robustness check. We present the implications for research and practice from two perspectives –managerial and readers⁵.

1.4.1. Managerial Perspective

As online reviews are gaining in popularity, the information within the reviews will become one of the major drivers for consumers' adoption of products and services. It is therefore imperative for businesses to extract actionable business intelligence from the

⁵ Reader refers to anyone who read online reviews e.g. potential customers, movie goers who wish to find out what others say about the movie etc.

vast amount of user-generated online reviews in order to gain their share of the market pie.

For companies, online consumer reviews provide some advantages over traditional survey. They are free to access, provide up-to-the-minute information and reviews are at the brand-model specific level. The results in this dissertation present directions in which firms can increase the marketability of their products. Using the constructed methodologies and leveraging on the massive amount of reviews related to their products, managers can integrate customer preferences and satisfaction data into the marketing / advertising process

Our large-scale comprehensive analyses on the impact of consumer reviews could assist firms in understanding how consumers use reviews in the decision making process. Properly utilized, the results found can be helpful in various aspects of business intelligence, ranging from market analysis to product planning and targeted advertising. As online reviews significantly affect consumers' purchase decisions, this dissertation provide constructed methodologies which can help firms to better predict the impact of reviews.

Online consumer review information can also be useful for identifying consumer preferences, finding out product defects and in correcting inadvertent mistakes. However, our study suggests that firms should be aware of, and should make adjustments in response to, biases that exist in online consumer reviews. Since online posting is self-reported by consumers, there may be non-random sample bias. Also, given the amount of manipulative activity going on in the online reviews forum, it is also possible that the reviews may be written by their competitors. Although our

results suggest that manipulation of review have a positive effect on sales, the cases of manipulation found in our data is large. If such an activity expands, the consequences may be detrimental to the credibility of online reviews.

1.4.2. Readers' Perspective

The findings in this dissertation show that consumers' rating behaviors are affected by *cultural influences*. Based on the data collected from IMDB.com and DOUBAN.com, we found significant differences across raters from these two different cultures. As user-generated online reviews are becoming an important source of information for consumers to make purchase or investment decisions, there is a need to understand that online reviewing behavior differs greatly from market to market, and might cause a reader to misjudge the quality of the product.

Also, online consumer reviews are *subjected to under-reporting bias*, which is, consumers with extreme opinions are more likely to report their opinions than consumers with moderate reviews causing online reviews to be a biased estimator of a product's true quality. Hence, we compare the consumer reviews posted online with those from an experimental study. Our results shows that under-reporting is more prevalent among U.S. online movie reviews, thus online movie reviews are a better perceived quality proxy in China and Singapore than in the U.S. If online behavior is not representative of offline behavior, and if the differences between online and offline behavior vary by nation, readers have to be aware that cultural influences and bias exists in online reviews and therefore should adjust their judgments accordingly.

1.5. Organization of the Thesis

The remainder of the dissertation is organized as follows. Chapter 2 is a literature review, which examines relevant research in marketing and text mining literatures that have studied the association between sales and reviews. Chapter 3 provides the dataset description and methodology for sentiment extraction and mining. Chapter 4 studies the economic impact of ratings, sentiments and their differential impact on sales. Chapter 5 examines the issue of review manipulation, the method to detect manipulation and the influence of such manipulation. Chapter 6 shed light on the existence of cultural differences in review posting and the existence of under-reporting bias. Finally, Chapter 7 concludes the dissertation with a review of the main results and an agenda for future research, with some preliminary findings.

Chapter 2. Background and Related Work

UGC research covers a broad range of topics and has fueled interest and enthusiasm from information systems scientist, computational linguist to marketing scientists and psychologists alike. In this chapter, we discuss some of the background and related work in the scope of our primary question: “What is the value of online reviews?” To estimate the value of online reviews, we explore some of the techniques used by marketing and text mining researchers.

Section 2.1 covers the related work on UGC in marketing which mostly examines the quantitative information such as ratings and volume of reviews. Section 2.2 covers the related work on sentiment mining and its challenges in automation. Finally, Section 2.3 presents the issues which current research have not considered and how the studies in this dissertation attempt to tackle them.

2.1. Word of Mouth

2.1.1. Quantitative Analysis – A Marketing Perspective

This section covers the related work which mainly examines the quantitative information of the online review. A majority of the work are largely drawn from the marketing literature with a couple from information systems literature.

Prior literature in marketing mainly focused on measuring the impact on product sales from two dimensions of online reviews, (1) the *volume* and (2) the *valence* (e.g. Liu 2006, Zhang et al. 2004). Volume measures the number of online reviews, and has been used to see the impact of online reviews on product sales (e.g. Chevalier and Mayzlin 2006). Therefore, the high volume of product reviews increases the promotion of the product and thus generates high product sales (Liu 2006). Valence on the other hand, measures the positive or negative opinions of online reviews (either based on the ratings or human coded as in Reddy et al. 1998 and Liu 2006). Unlike volume, the impact from the valence of online reviews is mixed. For example, using user reviews on Yahoo! Movies, Liu (2006) and Duan et al. (2008) found that the valence of previous movie reviews does not have significant impact on later weekly box office revenues. Zhang and Dellarocas (2006) on the other hand found a significant relationship between the valence of online WOM and box office revenues.

Apart from looking at the impact of online reviews, some researchers began to consider the pattern of reviews. For example, Hu et al. (2006) examined the aggregate pattern of online reviews and found that online reviews ratings reveal either a U or J-shaped pattern. They showed that most online reviews are either extremely positive (e.g. 5 stars in a 5-star review system) or extremely negative (e.g. 1 star). Few

reviews have moderate ratings (e.g. 3 stars). Duan et al. (2008) characterize the process of the feedback mechanism between word-of-mouth and retail sales through a dynamic simultaneous equation system. They show that a movie's box office revenue and word-of-mouth positive valence significantly influence word-of-mouth volume which in turn leads to higher box office performance. Moe and Trusov (2010) model the arrival of posted product ratings to measure the impact of social dynamics that may occur in the ratings environment and on both subsequent rating behavior as well as product sales. Godes and Silva (2006) investigate the evolution of ratings over time and order. They argue that the more ratings there are, the more dissimilar a shopper is from the entire set of previous reviewers and this leads to more purchase errors and thus lower ratings. Li and Hitt (2008) compared the early reviews with late reviews and tried to identify the difference in ratings between reviews at different time window. They argued that due to consumer heterogeneity and self-selection bias, early reviews could be systematically different from late reviews which may deliver biased opinions on the product. They reported evidence showing that for some books, early review ratings could be systematically higher or lower than the late reviews. Thus, they concluded that early review bias exists and could potentially reduce future consumer surplus.

Several other researchers have also actively examined the various effects of WOM (e.g. Chintagunta 2010; Moe and Trusov 2009; Godes and Silva 2009; Duan, Gu and Whinston 2008; Dellarocas, Zhang and Awad 2008; Li and Hitt 2008; Liu 2006; Clemons, Gao and Hitt 2006; Dellarocas and Narayan 2006; Chevalier and Mayzlin 2006; Hu et al. 2006; Godes and Mayzlin 2004; Dellarocas et al. 2004), primarily

analyzing the quantitative and numeric aspects of product reviews while ignoring the unstructured text comments in the reviews. Gruhl et al. (2005) show that volume of blog postings can be used to predict spikes in actual consumer purchase decisions at online retailer Amazon. Other researchers started to investigate various factors that could influence online reviews such as the impact of online reviewers' characteristics (Forman et al 2008; Ghose and Ipeiritis 2010) and product prices (Li and Hitt 2010). Forman et al. (2008) considers the effect of reviewers' online identities on the impact of reviews. They find that reviews posted by real name reviewers will have a larger impact on product sales than those posted by anonymous reviewers. Li and Hitt (2010) model the price effects in the reviews and suggest that companies should consider such effects when developing optimal pricing strategies.

However, these WOM research has focused on analyzing numeric ratings, mostly ignoring vast amounts of qualitative text product reviews. To fill this research gap, we intend to determine whether we can elicit additional useful information from qualitative text reviews beyond what we can learn from quantitative ratings.

As Chevalier and Mayzlin (2006) have indicated that consumers read text reviews rather than relying on only summary numeric ratings, one major challenge of research in this area would be to find ways to analyze the vast amounts of unstructured qualitative information.

2.2. Mining Online Reviews

2.2.1. Qualitative Analysis – A Sentiment Mining Perspective

As the volume of online review expands, it poses difficulty in finding and monitoring the relevant sources. Thus, automated sentiment analysis becomes a need. It is however, a challenging natural language processing / text mining problem.

The primary way to represent a document is by using the bag-of-word model: A document is represented entirely by the words that it contains and how many times the word appears, neglecting the order in which the words appear. This representation, although rather simplistic, delivers surprisingly good results in performing a diverse array of tasks. It is also one of the most viable approaches in processing millions of documents' sentiments.

Other sentiment mining techniques employed include combinations of machine learning, natural language processing methods and bags-of-words approach (Dave et al. 2003, Liu et al. 2005, Pang et al. 2002, Turney 2002). Previous work on sentiment analysis uses automatically generated sentiment lexicons, in which a list of seed words is used to determine whether a sentence contains positive or negative sentiments. Then, the polarity (i.e. positive or negative direction) of an opinion is identified based on the words within the review.

A simple machine learning approach for classifying products and services as recommended (thumbs up) or not recommended (thumbs down) was proposed by Turney (2002). Another approach for semantic classification of product reviews was presented in Dave et al. (2003). It involves a feature mining technique that is used to identify product features e.g. digital products such as camera. Thereafter, sentences

that give positive or negative sentiments for a product feature (e.g. picture quality, size, lens etc.) are extracted to give a summary on the comments of the opinions.

Apart from simply classifying the sentiments expressed as binary variables indicating that they are either negative or positive, greater depth in extracting richer, contextual sentiments and relating those to product choice or sales is needed. However, even the two-category version of the rating-inference problem for movie reviews has proven quite challenging for many automated classification techniques (Pang et al 2002, Turney 2002). Nevertheless, Pang et al 2005 has addressed the rating-inference problem in determining an author's evaluation with respect to a multi-point scale. Based on a supervised learning experiment, they achieved about 54.6% in accuracy for a four-class labeled data. Their technique however, requires annotated data and does not consider objective sentences.

Recently, the application of text mining techniques on online product reviews has also begun to draw the attention of text mining and information systems management researchers. Table 2.2.1 presents some selected relevant research that has applied sentiment analysis and linked online word-of-mouth to sales.

Table 2.2.1: Selected Studies in Word-of-Mouth Research

	Rating	Sentiments		Dependent variable	Interaction between ratings & sentiments	Impact of segments	Sample size (context)	If text is used, how many reviews analyzed?
		Manual coding	Text Mining					
Archak et al (2007)	√		√	Sales Rank			115 (Camera & Photo); 127 (Audio & Video)	1,955 Camera & Photo reviews; 2,580 Audio and Video Reviews
Berger et al. (2010)			√	Sales Quantity			244 (Books)	1942 reviews
Chevalier & Mayzlin (2006)	√			Sales Rank			2,387 (Books)	-
Chintagunta et al. (2010)	√			Box Office Sales			148 (Movies)	-
Clemons (2006)	√			Beer Sales			1,159 (Beer companies)	-
Das & Chen (2007)			√	Stock Prices			145,110 (Stocks)	145,110 messages
Dellarocas et al (2004)	√			Box Office Sales			80 (Movies)	-
Duan et al (2008)	√			Box Office Sales & Total reviews			71 (Movies)	-
Feldman et al. (2010)			√	Co-occurrence			135 (Car models); 5 (Drug forums)	868,174 messages (car); 671,102 messages (drug)
Forman et al (2008)	√			Sales Rank			786 (Books)	-
Fowdur et al. (2009)			√	Market Share			982 (Movies)	982 movie plots
Ghose & Ipeirotis (2008)	√		√	Sales Rank			144 (Audio & video players); 109 (Digital Cameras); 158 (DVDs)	522 reviews (Audio); 3,795 reviews (Digital camera); 431 reviews (DVD)
Godes & Mayzlin (2004)	√			Rating			44 (TV shows)	-
Li & Hitt (2008)	√			Sales Rank			2,651 (Books)	-
Liu (2006)	√	√		Box Office Sales			40 (Movies)	12,136 messages
Reddy et al (1998)			√	Box Office Sales & Longevity of Show			142 (Broadway shows)	1,254 reviews
Tsang & Prendergast (2009)	√	√		Interestingness, Trustworthiness & Purchase Intention	√		24 reviews (Controlled experiment)	-
Zhang & Dellarocas (2006)	√			Box Office Sales			128 (Movies)	-
This work	√		√	Sales Rank	√		50,380 (Books)	737,284 reviews

From Table 2.2.1, we find that one of the areas for potential investigation in the area of online reviews lies in the ability to analyze large-scale observations and extract relevant sentiment information from the reviews. Apart from simply classifying the sentiments expressed as binary variables indicating that they are either negative or positive, greater depth in extracting richer, contextual sentiments is needed. In the next section, we review the techniques for such automation.

2.2.2. Scale and the need for automation

“The meaning coded into words can’t be measured in bytes. It’s deeply compressed. Twelve words from Voltaire can hold a lifetime of experience.”

- Mark Horowitz, “Visualizing Big Data.”⁶

⁶ Wired, June 23, 2008, http://www.wired.com/science/discoveries/magazine/16-07/pb_visualizing

Sentiment mining is still technically challenging. There are still tough issues on human communication norms and computer processing limitations. In reality, it takes a person to understand a person – and even then it is easy to misunderstand sometimes.

Although, it is possible to build a statistical model using a sample of manually annotated documents and then automatically score the remaining documents, this may not truly represent the customer’s opinion – just the reader’s interpretation of what the customer thinks.

Suppose, we have rich human resources to manually annotate documents to determine if a statement is positive, negative or neutral. Even for human coders, this is not an easy task. Consider the following:

The cake is in the oven. – Neutral

The cake is delicious. – Positive

The cake is the worst I’ve ever tasted. – Negative

The cake is inedible. – Difficult to tell. It might just be a statement of fact.

The cake is better than my mother’s. – Sarcastic and with faint praise?

From this, we can see that scoring sentiments on a multi-point scale is even more difficult. In addition, sarcasm, irony, slangs are some nemeses of automated sentiment classification and this is only part of the problem. Sentiments can also be very different from conventional norm, for example, ‘disgusting and horrible’ can be bad when applied to food, but a good thing when applied to horror movies. Sadly, none of the sentiment mining techniques today are able to provide *scalable* and *effective* solutions to these problems.

Although the use of natural language processing to reveal the sentiment of reviews is still at the infancy, the techniques will improve in time. While not necessarily precise, the trends of sentiments revealed are certainly indicative and compelling for decision makers. Thus sentiment analysis should be viewed as a directional tool instead of a silver bullet. It is the trend that matters and they present directional managerial implications. Despite the imperfection of the techniques, it is still important to leverage the large amount of text reviews available and study their value and influence on consumers.

Given the vast amount of reviews available online, it is best to take advantage of the large-scale data so as to understand the sentiments of the massive online review population and their influence. To achieve this, we have to use a method that is viable on automatically extracting sentiments from consumer reviews for 1) a large number of reviews and is able to 2) quantify sentiments on a multi-point scale.

A few recent studies have applied various text mining techniques to quantify and analyze text product reviews, focusing on functional products represented by digital cameras (Archak et al. 2007; Ghose and Ipeirotis 2010; Hu and Liu 2004). Hu and Liu (2004) provide approaches for analyzing and comparing customer reviews and product reputation. Ghose and Ipeirotis (2010) perform analysis at the lexical, grammatical, semantic, and stylistic levels to identify text features that have high predictive power in identifying the perceived usefulness and the economic impact of a review. Furthermore, they examine whether the past history and characteristics of a reviewer can be a predictor for the usefulness and impact of a review. In this task, one essential issue is on how to quantify the qualitative reviews into quantifiable

information that can be integrated into a linear regression model (Archak et al. 2007; Ghose, and Ipeirotis 2010). Drawing upon their techniques, our research extends this line of effort to look into how ratings, sentiments and their interplay can influence the sale of books in Amazon.

Besides the various researches on the ratings and/or sentiments of online reviews, there are some aspects of online review to consider in assessing its value. The next section explores the gaps in current work and examines the characteristics, assumptions and bias of online reviews.

2.3. Gaps in Current Research for Online Reviews

2.3.1. Structure and Characteristics

In several online review websites such as Amazon.com, IMDB.com or Epinions⁷, the online review structure / design consists of the review *title*, which is a highlighted short summary of the review *content / body*. Since members may often read only some *titles* without reading detailed review *content*, reviews *titles* may have a differential impact on sales the than detailed review *bodies*. However, the format of online reviews has not been carefully considered in current research. Thus, in consideration of this design, we conduct sentiment mining on the review *title* and review *content* to gauge their differential impact.

In terms of online review characteristics, it has been found that online review ratings reveal the J-shaped distribution. This pattern stems from two self-selection biases – purchasing bias and under-reporting bias (Hu, Pavlou, and Zhang 2006).

⁷ The online reviews on these websites have been used by tremendous number of studies such as Hu et al. 2006, Duan, Gu and Whinston 2008, Li and Hitt 2008, Chevalier and Mayzlin 2006, Forman et al 2008, Archak et al. 2006, Ghose and Ipeirotis 2010 and so on.

First, consumers with higher product valuations who purchase a product have a higher tendency to write reviews than those with lower valuations are less likely to purchase (purchasing bias). Second, among consumers who purchased a product, those with the extreme ratings are more likely to express their views to “brag or moan” (under-reporting bias). This finding provides a backdrop on the modeling of our empirical analyses which current empirical models of UGC have not considered.

With the understanding of the design and characteristics of online review, we can take these into account for our empirical analyses and improve on the veracity of the results and this issue is addressed in Chapter 4. Also, with the extensive global usage of online reviews, it would be important to examine if the characteristics of reviews are consistent across cultures and we will be examining this in Chapter 6.

2.3.2. Assumptions and Bias

With the proliferation of online review systems, many people believe that online consumer reviews are a good proxy for overall word-of-mouth and can also influence consumers’ decisions. The efficacy of online reviews could nonetheless be limited. First, reviewers are not a randomly drawn sample of the user population. Anderson (1998) finds that extremely satisfied and extremely dissatisfied customers are more likely to initiate WOM transfers. This led to the findings by Hu et al. (2006) who found that online review ratings revealed J-shaped distribution characteristics; while Li and Hitt (2008) find potential bias in consumer reviews during early product introduction periods. Secondly, interested parties can easily manipulate online forums. Dellarocas (2006) and Mayzlin (2006) theoretically analyze scenarios in which firms

can anonymously post online reviews to praise their products or to increase awareness about them.

Given the power of electronic word-of-mouth, many firms are taking advantage of online consumer reviews as a new marketing tool (Dellarocas 2003). Studies show that firms not only regularly post their product information and sponsor promotional chats on online forums, such as USENET (Mayzlin 2006), they also proactively induce their consumers to spread the word about their products online (Godes and Mayzlin 2004). Some firms even strategically manipulate online reviews in an effort to influence consumers' purchase decisions (Dellarocas 2003; Harmon 2004). An underlying belief behind such strategies is that online consumer reviews can significantly influence consumers' purchasing decisions. Some recent studies have looked into how marketers can strategically manipulate consumers' online communications (Dellarocas 2003; Mayzlin 2006).

Manipulation or fraud is not a new area of research in the traditional business fields. For example, in the area of accounting there is extant research on the profiling of earnings manipulators through the identification of their distinguishing characteristics as well as the development of models to detect earnings management (Beneish 1999; Chevalier and Goolsbee 2003). The variables used in such models represented the effects of manipulation or preconditions that prompted firms to engage in such activities. Research in this area identified the existence of a systematic relationship between the probability of manipulation and some key financial statement variables. As a result, the analysis of the accounting data of the companies could be used to identify firms which have engaged in earnings manipulation. In fact,

by comparing the accrual levels for one company over different years and under different types of financial situations, the investigator is able to identify the abnormal accruals that were closely related to earnings management. Although the models used in the earnings manipulation literature were easy to implement, the financial reports of the same company had to be available for over several years to effectively detect such fraudulent activity.

Another similar area of research is the detection of Internet click fraud. A click fraudster is defined as a person or an automated computer program that imitated the online behavior of a legitimate user using a web browser. This is done by clicking on a web-based advertisement for the purpose of generating a charge per click without having any real interest in the content of the advertisement. Click fraud can be conducted using either automated robots or human agents. A common way to conduct this is by clicking on the advertisers' hypertext links that were displayed on websites or listed in the results of search engine queries using programmed robots. An alternative way to do this is -by hiring low-cost workers from developing countries to manually click on the advertisements (Majumdar 2007). Both methods can involve clicking advertisements on own websites to gain more revenue from advertisers who often made payments on the basis of the number of clicks that the advertisements received, or clicking on advertisements placed in websites of competitors to waste their marketing budget and skew search results⁸. Various methods are available for the detection of online click fraud. These methods generally use a combination of web traffic analysis, and web surfing behavior recognition. For example, Metwally,

⁸ http://news.cnet.com/Exposing-click-fraud/2100-1024_3-5273078.html

Agrawal and El Abbadi (2005) used association rules for the detection of fraud in web advertising networks, whereas Majumdar, Kulkarni, and Ravishankar (2007) proposed protocols that could be used to identify fraudulent behavior by brokers and other intermediaries in content-delivery networks.

However, the models used in the detection of earnings management or click fraud could not be adopted directly for the detection of online reviews manipulation due to some unique features of the online review environment, which present new challenges:

- 1) Manipulators in the online environment could assume any identity that resembled a real consumer, or could even remain anonymous;
- 2) Manipulators use both numeric ratings and text comments to influence potential consumers' purchase decisions; and
- 3) Techniques used for the detection of click fraud required knowledge of the IP addresses and having access to the proprietary users browsing behavior data that were generating the clicks, and such knowledge and data are usually available only to the online vendors. However, in case of online reviews the IP addresses of the reviewers were only available to the site administrators and could not be released to the public due to privacy concerns.

Due to the above challenges, a method for the detecting manipulation in online reviews is crucial. In Chapter 5, the objectives will be to develop a method for detecting manipulation and understand the impact of such activities.

To wrap up, this chapter has examined relevant work that examines the value of online reviews / WOM. Most of the studies however, either examine the effect of

quantitative information of online reviews such as the volume of reviews and numeric ratings or just the textual sentiments and both ratings and sentiments nor their interplay. Taking into consideration of the design of online reviews, a comprehensive and large-scale systematic study on the rating-sentiment influence is necessary. Also, the ability to detect manipulation and examine its impact on sales is another issue that needs to be addressed. Finally, for users of online reviews (be it firms or individuals), it is important to assess the cultural influence on online reviews reporting and assess the level of under-reporting bias across countries.

Chapter 3.

Methodology and Measurements

This chapter describes the system design for sentiment mining in the studies of this dissertation. The main contribution of this system is that it is viable on automatically extracting sentiments for 1) a large number of reviews and is able to 2) quantify sentiments on a multi-point scale. Leveraging large number of reviews allows us to understand the sentiments of the massive online review population and their influence, while the ability to quantify sentiments on a multi-point scale allows us to empirically estimate the impact of sentiments and sets the direction for managerial strategy.

The first section describes the system designs we have employed to perform our sentiment mining. Section 3.2 presents additional measurements necessary for understanding the writing style of each review which will be used in our technique for manipulation detection. Section 3.3 introduces the dataset used in the studies.

3.1. Technology & systems for scalable, semi-automated execution

Figure 3.1 shows a typical online review in Amazon. For each online review, we have the numeric average ratings and text review provided by a reviewer. Specifically for sentiments, we have two sub-components, sentiments in the title and sentiments in the content of the review. It is likely that the consumers might pay more attention to the sentiments in the titles of the review as 1) the titles summarize the overall sentiments of the customer; 3) it is highlighted in bold on each customer review leading to greater attention, and 2) processing the title information requires limited cognitive power and time. We intend to investigate the differential impact of sentiments in customer review title as well as the sentiments in the content on product sales. To do so, we have to first perform sentiment mining on the online reviews. The next subsection describes the systems for scalable and semi-automated execution of sentiment analysis.

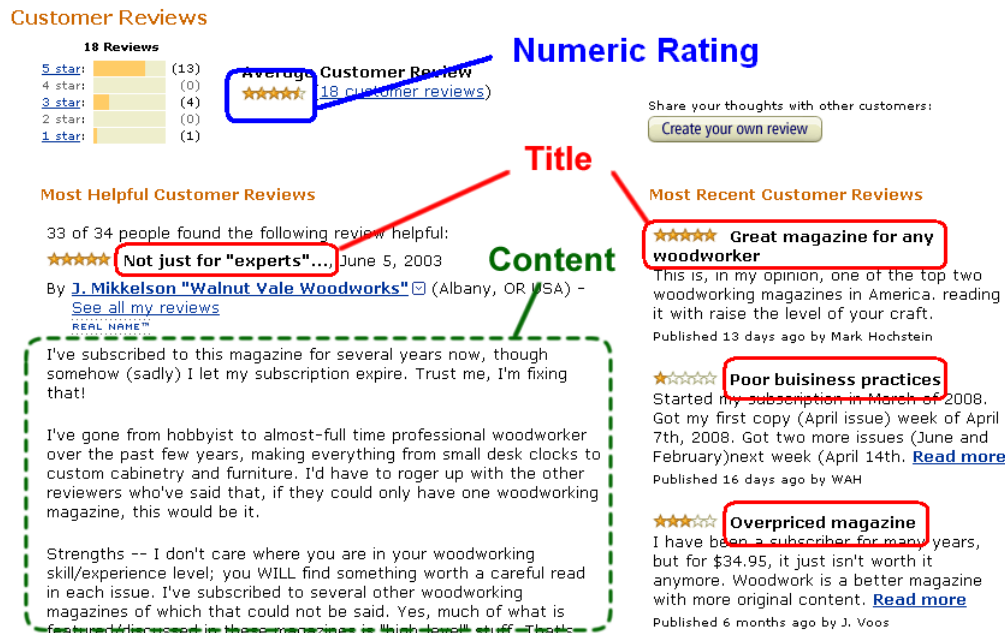


Figure 3.1: Screenshot of Customer Reviews in Amazon.com

3.1.1. System designs for sentiment extraction and mining

Instead of merely determining whether a review is “thumbs up” or not, we attempt to infer the reviewer’s implied numerical evaluation from the text sentiments on a multi-point scale. Figure 3.1.1a presents the main system design of our sentiment extraction and mining process.

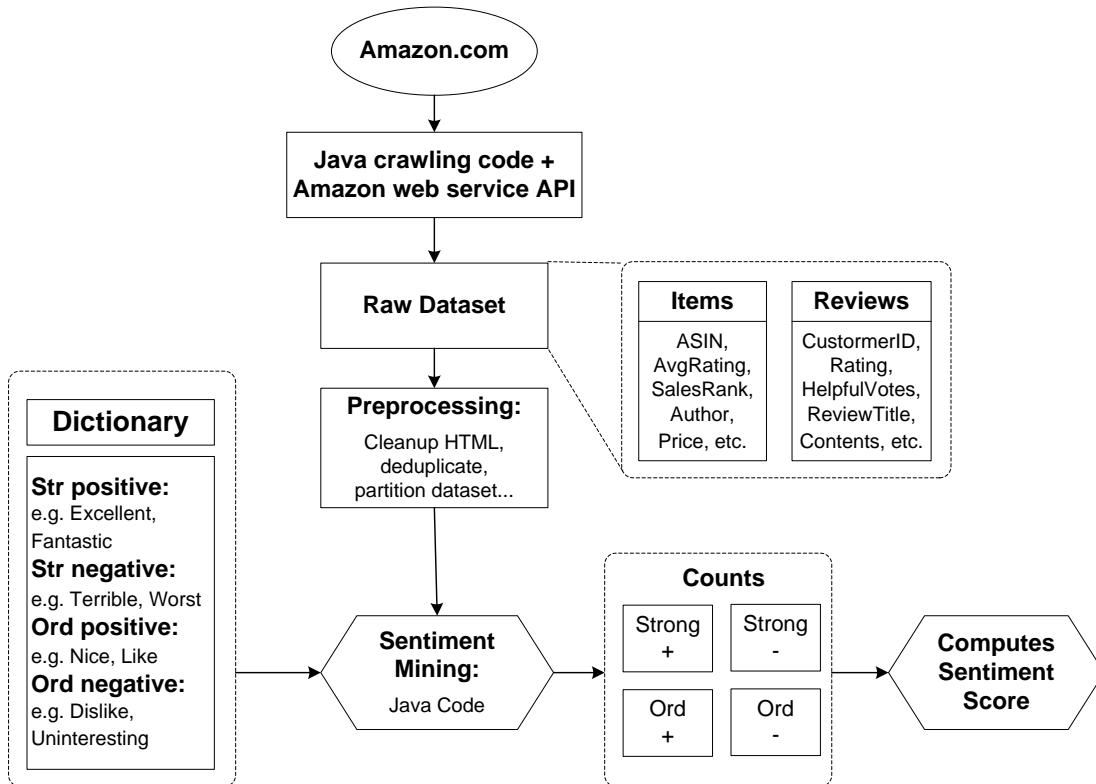


Figure 3.1.1a: System Design of Sentiment Extraction and Mining Process

The sequence of tasks is as follows. We wrote a java crawling code using the Amazon Web Service Application Programming Interface (API) to collect product information and online reviews from Amazon Web Service.⁹ Structure analysis is performed for each online review such that the numeric and text information are

⁹ The crawler is specifically designed for the crawling of Amazon review data, however it can also be easily customized to crawl reviews from other websites such as movie reviews from IMDB and Douban.com. Alternatively, an open source crawler named webscraper works just as well.

segmented for further processing. The dataset we obtained is from book category and there are two separate tables – book items and book reviews. The first table will contain book item information such as the unique ASIN identification number, the average rating, sales rank, and price for each item. The second table contains the customer review information such as customer ID, rating, helpful votes for each review. Due to the size of the dataset, we have to partition the data into several tables because of technical limitations and memory constraints. Subsequently, the text contents of each review were hashed for duplicate removal.

Dictionary Construction

To compile the sentiment word list, we used a manual approach combined with a dictionary-based approach. The system contains 2 supplementary sentiment databases that were used to support the sentiment scoring algorithm.

- First, an electronic General Inquirer Dictionary (Stone et al 1966)¹⁰, which provides the base language data where each word has been pre-tagged on its polarity, i.e. positive or negative.
- Second, a “lexicon” which is a manual-picked collection of strong positive and negative words that were found from the reviews of Amazon.com.¹¹ This will help to increase the accuracy of our sentiment scoring as these terms are commonly found in Amazon reviews.

¹⁰ The General Inquirer (GI) lexicon has been used by Hatzivassiloglou and Mckeown (1997), Turney and Littman (2002) and several others in their research work. The GI is a useful resource for content analysis of text. It consists of words drawn from several dictionaries and grouped into various semantic categories. It lists different senses of a term and for each sense it provides several tags indicating the different semantic categories it belongs to.

¹¹ The terms/phrases were obtained by Archak et al (2007) from the reviews in Amazon.com. Each term / phrase is given a score on the scale of 0 to 100. Among the 2697 terms/phrases they obtained, we extracted 40 strong positive terms (with scores higher than 95) and 30 strong negative terms (with scores less than 30).

The list of words from the dictionary formed the lists of ordinary sentiment terms while those in the lexicon form the lists of strong sentiment terms. The strong positive sentiments are terms like ‘excellent’ and ‘awesome’ which are commonly found in Amazon online reviews while strong negative sentiments are terms like ‘terrible’ and ‘awful’. The ordinary positive terms (e.g. nice, satisfactory) and ordinary negative terms (e.g. redundant, dislike) were collected from a publicly available online dictionary where each word has been pre-tagged as either positive or negative.

Based on these 4 lists of seed words, we perform lexicon expansion and then calculate the number of sentiment terms in each review to obtain the sentiment score. Lexicon expansion is performed by finding the various morphological forms of words from the list of seed words in the dictionary and the lexicon e.g. like and likes are different forms of the same lexeme.

Although the construction and choice of dictionaries based on the manual approach were specialized to our dataset, the process itself is a general method that has been used in current work (Liu 2010). As the manual approach is very time-consuming, it is therefore usually combined with automated approaches such as the dictionary-based approach as the final check because automated methods make mistakes.

Sentiment Analysis

Sentiment (or polarity) analysis is then performed to identify positive and negative language in text. In our system, we used a general approach that is scalable for our size of data in which the polarity and strength of an opinion is estimated based on the occurrences of sentiment words within the title and the content (Archak et al.

2007, Das and Chen 2007). We gave a higher weight of 2 and -2 for the strong positive and strong negative terms respectively. The ordinary positive term is given a weight of 1 and ordinary negative term is given a weight of -1. Each word in the title and content is checked against the sentiment database and assigned a count value (± 1 , ± 2). The output is a computed sentiment score for each review.

Thus, for each review, we compute the number of sentiment term occurrences within a review. The polarity and strength of an opinion is calculated based on the occurrences of the sentiment words times their individual weights within the review. The difference between the positive terms and negative terms are normalized by the total number of sentiment terms to discount the influence of longer reviews. The sentiment score of a customer review i is computed as:

$$senti_score_i = \frac{(str_pos_i * wg + ord_pos_i) - (str_neg_i * wg + ord_neg_i)}{(str_pos_i + str_neg_i) * wg + ord_pos_i + ord_neg_i} \quad (3.1a)$$

where:

- str_pos_i : the number of strong positive terms in review i
- str_neg_i : the number of strong negative terms in review i
- ord_pos_i : the number of ordinary positive terms in review i
- ord_neg_i : the number of ordinary negative terms in review i
- wg : the weight of strong terms

Then from Equation 3.1a, the minimum and maximum sentiment score obtained from each review will be $[-1, 1]$. For product item j , there are n reviews. Within the text of each review, it contains a title and the content. Thus, for the i^{th} review of product j , we first compute its sentiment score for the title and the content separately

using Equation 3.1a. Then the sentiment of the i^{th} review is: $(title_score_i + content_score_i)/2$. To facilitate comparisons between the numeric rating and the sentiment score, we convert the sentiment score for each review to a scale of 1 to 5, rounding to one decimal point, which is similar to that of the average numeric rating scale in Amazon.com.¹² We have tried other alternative measures such as percentage count of the sentiment terms and the difference of average number of strong positive and negative sentiments. We chose to use the measure in Equation 3.1a as it is comprehensive in capturing the essence of all types of sentiments and is able to provide a quantifiable and comparable sentiment score that can be manually judged.

Two judges were recruited and were provided the task to independently read a sample of 200 reviews. For each review, the judges were asked to gauge if the sentiment score is accurate and reflects the overall sentiments expressed in the review. The reviews were randomized and both judges rated the reviews in the same order to avoid order biases. We conducted inter-judge reliability tests to determine the extent of agreement shown by the two judges in assessing the proper reflection of the sentiment score in the sample of reviews. Overall, there is significant agreement between the two judges on the sentiment score of all reviews with Cohen's kappa of 0.8521.

¹² Suppose from Equation 3.1a, we obtain a sentiment score of $x = 0$, the sentiment score on a 1 to 5 scale is:
 $f(x) = 2x + 3 = 3$.

Robustness Check

To ensure that our empirical results obtained from our sentiment scores are robust, we have performed another set of empirical analyses using sentiment scores obtained from a more complex but less scalable system.¹³ In this system, the weight for each term is determined through training and tuning as depicted in Figure 3.1.1b.

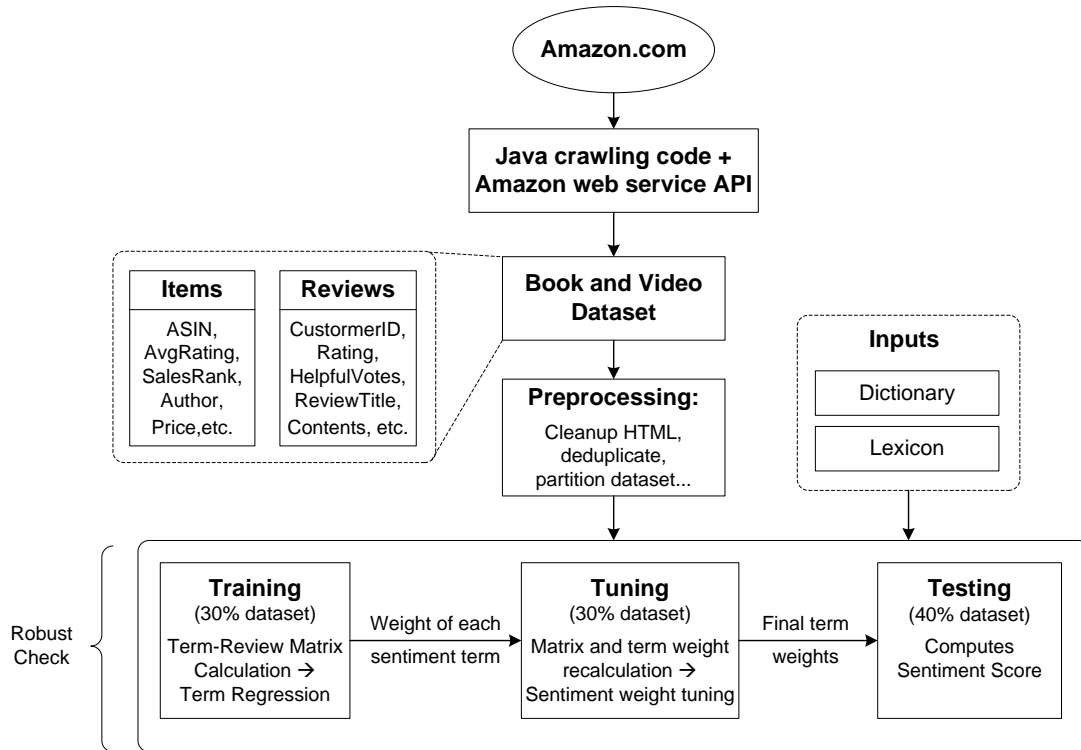


Figure 3.1.1b: System Design of Sentiment Extraction and Mining Process

For the robustness check, instead of grouping the sentiment terms into either strong or ordinary types, we estimate a weight for each individual sentiment term. In this system, we draw upon the work of Archak et al. (2007) to check the robustness of the empirical results obtained by our prior system. Likewise, the list of words from the dictionary and those manually extracted will form a list of sentiment terms in this system. Based on this

¹³ The empirical results are qualitatively similar for both systems and the results are presented in Chapter 4.

list of seed words, we proceed to determine the weight for each of the sentiment term. Figure 3.1.1c shows how the preprocessed dataset is partition for the term weight calibration process.

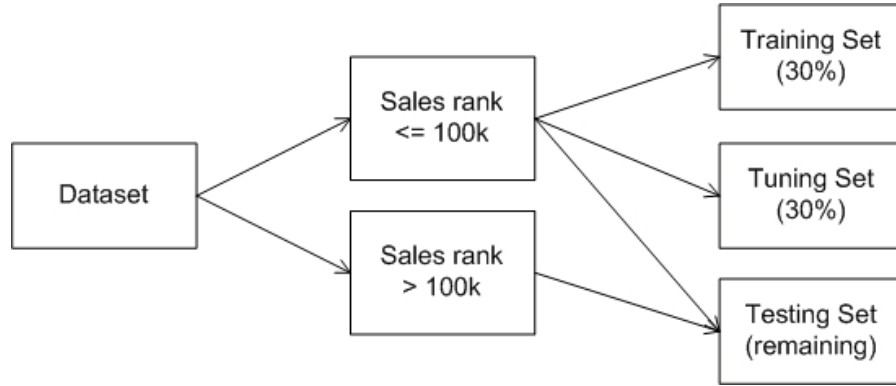


Figure 3.1.1c: Term Weight Calibration Process

Term \ Review	Review												...
	1	2	3	4	5	6	7	8	9	10	11	12	
Good	0	0	2	0	1	1	0	0	2	0	0	2	...
Like	1	1	0	0	0	2	0	0	0	3	0	0	...
Bad	0	1	0	1	0	0	0	2	0	0	2	0	...
Terrible	0	0	0	1	0	0	0	0	0	1	0	0	...
...													

Figure 3.1.1d: Term-Review Matrix Generated

The calibration process is as follows:

Step 1: For book dataset D , extract items with sales rank $\leq 100k$

Step 2: From Step 1, randomly select 30% as the training set (D_{train}).

Step 3: For D_{train} , count the sentiment word occurrence in each review. This gives us a term-review matrix (see Figure 3.1.1d). Sum the value to the item level such that for each ASIN, we have the frequency occurrence for each sentiment term. ($N_{\text{ASIN},i}$, where i is the index of the term).

Step 4: Normalize the value of $N_{ASIN, i}$, for each i , we obtain the $Max_i = \max(N_{ASIN, i})$ for all book items. Finally for each ASIN, the normalized of $N_{ASIN, i} = N_{ASIN, i} / Max_i$. From this, the final term-review matrix obtained for all the sentiment terms are placed on a scale of 0 to 1.

Step 5: Using the normalized $N_{ASIN, i}$, we perform regression to get the weight of each sentiment term. The coefficient of each sentiment term represents the weight.

Step 6: Select another 30% as the tuning set from Step 1. For the tuning set (D_{tune}), repeat the procedure in step3 to 5.

Step 7: Using the term weights, calculate the new sentiment score for the test set (D_{test}).

In this robust check, we estimate a different set of weights for the sentiment terms. First, we extract product items that have sales rank from 1 to 100,000. This is because sales rank within this range is updated daily and are therefore more accurate than items with sales rank above 100,000.¹⁴ Then from these extracted items, 30% of the dataset is used to perform the training (training set), the other 30% for tuning the parameters (tuning set) and the remaining of the dataset for testing (test set). For the training set, we calculate the frequency of sentiment term occurrences for each review. The next step is a summation of these term frequencies to the product item level. The final term-review matrix is normalized to a scale of 0 to 1. Then, we estimate the weight of each sentiment

¹⁴ We have also tried the training and tuning procedure by randomly selecting items from the whole dataset without the sales rank classification. The empirical results obtained are qualitatively similar and the conclusions do not change. The results show that the sentiment has a much greater correlation on sales rank than numeric ratings; the sentiments in the content are much more impactful than those in the title; and ordinary sentiments are much more impactful than strong sentiments.

term using on the following regression model by controlling the price, total reviews and average rating for each product:

$$\ln(\text{SalesRank}) = \alpha_1 \ln(\text{Price}) + \alpha_2 \ln(\text{TotalReviews}) + \alpha_3 (\text{AvgRating}) + \alpha_4 \text{sentiment_term}_1 + \dots + \alpha_{n+4} \text{sentiment_term}_n + \varepsilon$$

The same procedure is carried out for the tuning set. Finally, the parameter estimate of each sentiment term from the training set and the tuning set are averaged to derive the final weight for each sentiment term. Sentiment terms with inconsistent polarity i.e. negative or positive are dropped. For each review, the sentiment score is calculated by the weighted sum of the number of positive terms minus the number of negative terms:

$$\frac{\sum_{i=1}^n \text{positive_term}_i * \text{wg}_i - \sum_{i=1}^m \text{negative_term}_i * \text{wg}_i}{\sum_{i=1}^n \text{positive_term}_i * \text{wg}_i + \sum_{i=1}^m \text{negative_term}_i * \text{wg}_i} \quad (3.1b)$$

where:

- positive_term: the number of positive terms
- negative_term: the number of negative terms
- wg: the weight of each sentiment term

The sentiment score derived from Equation 3.1b is then used in our robustness check of the empirical results which will be further elaborated in Chapter 4, Section 4.3.1.

3.1.2. Sentiment measurements

Once we derived the sentiment score for each review, our next step is to obtain an average sentiment score for each product item. The simple mean (Equation 3.2) is

used because it is the easiest proxy for inferring product sentiments, thus it is used as a potential predictor of product sales. For an item containing n reviews, the final average sentiment score for product j is:

$$AvgSentiScore = \left(\sum_{i=1}^n senti_score_i \right) / n \quad (3.2)$$

For robust checking purpose, we propose a weighted sentiment measure i.e. mean weighted by the percentage of consumers who think the review is helpful (Equation 3.3) based on the relative relationship among helpful votes (# customers think one review is useful) and total votes (# customers read that review). The empirical results using weighted average method were qualitatively similar to those using the simple average model. Also, to address the issue of J-shaped distribution of review ratings (Hu et al. 2006), we have checked the robustness of our results using beta distribution models. Using the mean and variance derived from the beta distribution¹⁵ shape parameters, we obtained qualitatively similar results to those using simple average model. Thus, in the interest of parsimony, we present the results using simple average.

$$AvgSentiScore(Helpfulvotes) = \frac{\sum_{i=1}^n factor_i * senti_score_i}{\sum_{i=1}^n factor_i} \quad (3.3)$$

Below, we define the different aspects of sentiment score used in this study:

¹⁵ We have tried modeling the pattern of the ratings and sentiments for each product item i using the beta distribution. The beta distribution is parameterized by two shape parameters, denoted by α and β . From the shape parameters estimated α and β , we conduct the regressions based on the mean and variance derived from the beta distribution shape parameters where $E(X) = \alpha / (\alpha + \beta)$ and $Var(X) = \alpha \beta / (\alpha + \beta)^2 (\alpha + \beta + 1)$. Results are qualitatively similar to those using simple average.

Average Title Sentiment Score

This is the sentiment score for each product item in which Equation 3.1 is applied on the title of the text review.

Average Content Sentiment Score

This is the sentiment score for each product item in which Equation 3.1 is applied on the content of the text review.

Average (Rating*Sentiment Score)

This term captures the interaction effect between the numerical ratings of a review and how that review was written. It represents the indirect impact of sentiment on sales through rating. A significant positive coefficient between this interaction term and sales means that reviews written with strong sentiments and with high ratings have a bigger impact on sales than reviews written with ordinary sentiments. To measure this interaction effect, for every review of product j , we multiply the rating with the sentiment score.¹⁶ Then the final interaction term for product j is:

$$\left(\sum_{i=1}^n rating_i * senti_score_i \right) / n \quad (3.4)$$

Average Strong Positive/Negative Score or Average Ordinary Positive/Negative Score

The strong positive/negative score or ordinary positive/negative score for the i^{th} review is calculated using the following formula:

¹⁶ The qualitative nature of the empirical results do not change if this interaction equation is the average rating of product j multiply by the average sentiment score of product j .

$$\frac{senti_part_i}{str_pos_i + str_neg_i + ord_pos_i + ord_neg_i},$$

where $senti_part_i \in \{str_pos_i, str_neg_i, ord_pos_i, ord_neg_i\}$

(3.5)

The strong positive/negative score or ordinary positive/negative score for each product is obtained from the content of each review. The final strong positive/negative score or ordinary positive/negative scores of a product is the average of all the strong positive/negative score or ordinary positive/negative scores over all the reviews received by that product item respectively.

3.2. Other measures on the stylistic content

As our second study examines online review manipulation, this section introduces other measures which are used in the detection of manipulated reviews.

Writing style refers to how consumers construct sentences together and it varies with the background of an individual. Intuitively, reviews written by different consumers will be random in the case of no review manipulation. Thus, by observing the change in the writing style across the reviews over time, we can infer whether the online reviews for a product is manipulated or not because writing style is unique amongst individuals. Building on this intuition, this section presents additional measurements to determine the writing style of a review which will be used in our model for manipulation detection in Chapter 5.

3.2.1. Stylistic measurements for manipulation detection

The readability of the reviews or the reader's ability to comprehend a text is measured using the Automated Readability Index (ARI) (Senter and Smith 1967).

Past research in the field of information science made use of readability tests for studying the qualitative characteristics of texts content (Ghose and Ipeirotis 2010; Paasche-Orlow et al. 2003). The ARI is one of the major readability tests used to evaluate the readability of a text by decomposing the text into its basic structural elements. We chose this measure because unlike other indices, the determination of ARI relied on the number of characters per word, rather than the number of syllables per word. Since, the number of characters in a word could be more easily and accurately determined than the number of syllables per word, this measure is subjected to less error as compared to other readability measures. The ARI is calculated using the following formula:

$$ARI = 4.71 (Total\ number\ of\ characters / Total\ number\ of\ words) + 0.5 (Total\ number\ of\ words / Total\ number\ of\ sentences) - 21.43 \tag{3.6}$$

The value of the index approximated the minimum grade level of education that was needed to comprehend a piece of text. For instance, a score of 8.3 for the ARI for a piece of text indicated that the text could be understood by an average 8th grade student in the United States.

3.3. Dataset Description

The focus of this dissertation is on single-purchase products. Information goods, such as books, movies, music, and computer games, are examples of products purchased only once. As many of these single-purchase products are considered experience goods (Nelson 1970), their product characteristics are difficult to observe

until consumption. Thus, online reviews are valuable in reducing the risk of purchasing such products and our interest is to examine the degree of such influence.

There are three main sources of our dataset – Amazon.com, IMDB.com and Douban.com. The data gathered from Amazon will help to address the first question on the estimating the value of online reviews while the data gathered from IMDB.com and Douban.com will help to address the second question on the informative worth of online reviews. Since textual analysis is only performed on the Amazon dataset, I will introduce only the Amazon data in this chapter. The text analysis metrics derived will then be used for our empirical analysis in Chapter 4 and 5. For ease of readability and understanding, I have moved the description of IMDB and Douban data to Chapter 6 since no textual processing have been performed on these two data.

3.3.1. Amazon cross-sectional dataset

Our data were gathered from Amazon using its Web Service (AWS) in August 2005¹⁷. We select Amazon as it is the leading electronic retailer for books representing 70% of the whole market transactions (Ehrens and Markus 2000). It has also been used to study research questions regarding online reviews by various previous studies (e.g. Chevalier and Mayzlin 2006, Hu et al. 2006, Archak et al 2007, David and Pinch 2005, Forman et al. 2008, Ghose and Ipeirotis 2010, Li and Hitt 2008). For each item, we collected the title, price, sales, and review information. Specifically, for each customer review, we gathered the numeric rating, review date, helpful votes, total votes and the original text. The text is separated into two parts:

¹⁷ The data are collected from 12/8/05 to 29/8/05.

title and content. The numeric ratings for each review are on a one-star to five-star scale where one-star corresponds to least satisfied/liked/preferred and five-star corresponds to most satisfied/liked/preferred. Amazon does not report the actual sales for the products; instead it provides a sales rank figure for each product which ranks the demand for a product relative to other products in the same category. Henceforth, product sales rank is shown in descending order where 1 represents the best selling product. Consequently, there is a negative correlation between product sales and sales rank. Prior research in economics and in marketing demonstrated that the distribution of demand in terms of sales rank has a Pareto distribution (Chevalier and Goolsbee 2003). Therefore, the Pareto relationship allows us to convert sales rank into demand levels and we can use log of product sales rank as a proxy for product sales. The summary statistics and descriptive statistics of our dataset are shown in Table 3.3.1a and Table 3.3.1b respectively.

Figure 3.3.1 and Figure 3.3.2 presents the exploratory data visualizations of the rating distributions and sentiment distributions respectively. Both distributions shows a trend of overwhelmingly positive reviews in Amazon and this is consistent with the trend found in other Amazon dataset such as those gathered by Chevalier and Mayzlin (2006). In particular, the rating distribution exhibits a J-shaped distribution as described by Hu et al (2006) and will be considered in our modeling approaches.

Table 3.3.1a: Summary Statistics of Dataset from Amazon.com

Product Category	Number of Products	Number of Online Product Reviews
Books	50,373	737,284

Table 3.3.1b: Descriptive Statistics of Books

Variable	Median	Mean (SD)
Retail Price	16.47	29.39 (30.59)
Sales Rank	122,103	204,331 (243,085)
Age of Product (days)	1,685	2,414 (2,553)
Average Rating	4.5	4.35 (0.73)
Number of Reviews	4	15.74 (76.10)
Average helpful ratio	0.83	0.79 (0.20)

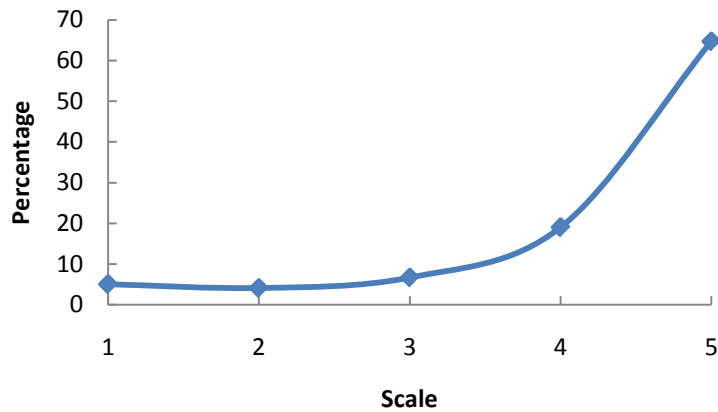


Figure 3.3.1: Rating Distribution

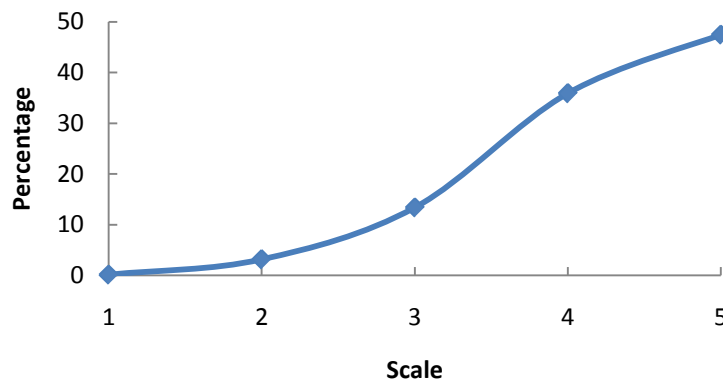


Figure 3.3.2: Sentiment Distribution

In this chapter, we have introduced the system design of our sentiment mining technique that is capable of scoring sentiments on a multi-point scale for large

datasets. In the next chapter, the metrics derived from the system are integrated into a regression models to estimate the value of online reviews. The first study examines how rating, sentiments and their interplay in online reviews affect sales when consumers turn to the “wisdom of the crowds” for decision making.

Chapter 4.

Impact of Sentiments on Sales

It is generally assumed that ratings are a numeric representation of text sentiments and their valences are consistent, this however may not always be true. Instead of investigating the two important elements of a review separately – rating and sentiments, we analyze how ratings, sentiments and their interplay affects consumers' purchase decision.

In this chapter, we conduct a systematic analysis to examine the value of online reviews for firms. Using a large scale cross-sectional data, with close to three quarter million online consumer reviews spanning over 10 years for over 50,000 books sold on Amazon.com, the results present compelling directional managerial and marketing implications.

4.1. Conceptual framework

As it is unlikely that consumers will view all reviews, our interest is to understand how consumers search, encode and abstract the reviews that are presented. Hence, we interviewed a couple of Amazon shoppers and conduct an experiential survey where participants were asked to rate the relative importance of ratings and sentiments at each stage of the decision making process.

From our in-depth interviews with some Amazon shoppers, Figure 4.1a offers one possible way that customers search and utilize the ratings and customer reviews in making their choices. This is particularly typical to the scenario when a customer does not have a particular book title in mind but only a topic and set of keywords of the kind of book he or she is interested in. The online search and decision making process is as follows: On the Amazon search space, a potential customer keys in a set of keywords and the search result returns a list of books (see example in Figure 4.1b) related to the keywords queried by the customer. Often, the search results returned may be in hundreds or even thousands and it is impossible for customers to evaluate all available alternatives in great depth. Thus, there are several stages in the decision making process before the consumers reach their final purchase decision.

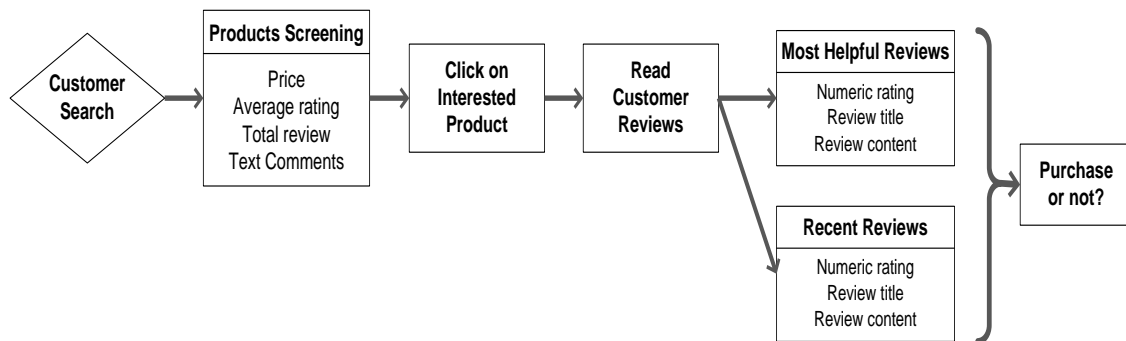




Figure 4.1a: General Framework of Customer Search and Decision Making Process

At the first stage, consumers typically screen a large set of relevant books and search for information to gain awareness on the lists of relevant books available. For each book item on the list, there will be information such as the price, average rating, total number of reviews, excerpts of text sentiments and so on. Using this information, consumers identify a subset to evaluate in greater depth. At the initial stage, the potential customer tends to use ratings to decide which book to click on and evaluate further. And generally, customers tend to click on and further evaluate book items with high average ratings.


Showing 1 - 16 of 826 Results Choose a Department to enable sorting

- 


New Zealand (Eyewitness Travel Guides) by Kate Hemphill (**Paperback** - Mar 15, 2010)
 Buy new: ~~\$26.00~~ **\$16.50**
 27 new from \$13.89 12 used from \$13.13
 Get it by **Thursday, Oct 21** if you order in the next **16 hours** and choose one-day shipping.
 Eligible for **FREE** Super Saver Shipping.
 ★★☆☆☆ (1)
Books: See all 808 items

- 

New Zealand (Country Guide) by Charles Rawlings-Way, Brett Atkinson, Sarah Bennett, and Peter Dragicevich (**Paperback** - Oct 1, 2010)
 Buy new: ~~\$26.00~~ **\$19.43**
 33 new from \$16.98 9 used from \$16.84
 Get it by **Thursday, Oct 21** if you order in the next **16 hours** and choose one-day shipping.
 Eligible for **FREE** Super Saver Shipping.
 ★★★★★ (35)
Books: See all 808 items

- 

Frommer's New Zealand (Frommer's Complete) by Adrienne Rewi (**Paperback** - Feb 15, 2010)
 Buy new: ~~\$22.00~~ **\$15.63**
 37 new from \$12.64 10 used from \$12.58
 Get it by **Thursday, Oct 21** if you order in the next **16 hours** and choose one-day shipping.
 Eligible for **FREE** Super Saver Shipping.
 ★★★★★ (23)
Excerpt - Copyright: "... she is based in Christchurch writing for numerous **New Zealand** and international magazines. She is the author of five editions of the bestselling **travel guide** *Frommer's New Zealand* and has published three other non-fiction titles ..."
Surprise me! See a random page in this book.
Books: See all 808 items

- 

The Rough Guide to New Zealand (Rough Guides) by Paul Whitfield (**Paperback** - Oct 4, 2010)
 Buy new: ~~\$27.00~~ **\$18.47**
 40 new from \$16.09 10 used from \$16.09
 Get it by **Thursday, Oct 21** if you order in the next **16 hours** and choose one-day shipping.
 Eligible for **FREE** Super Saver Shipping.
 ★★★★★ (3)
Books: See all 808 items

Figure 4.1b: An example of search results returned

When a customer clicks on a book and enters a particular book's webpage, the online reviews related to that book will be presented. By default, consumers will see two types of reviews – Most Helpful reviews and Most Recent reviews (see Figure 4.1c). Both types of reviews will have a mix of positive and negative reviews and consumers tend to

read both positive and negative reviews to get a balance view. Using these reviews, consumers make their evaluation and final choice. The feedback we obtained from the interviewees is the sentiments expressed in the reviews have a major influence on their purchase decision; the average rating for the book merely acts as an initial filter at the search stage.

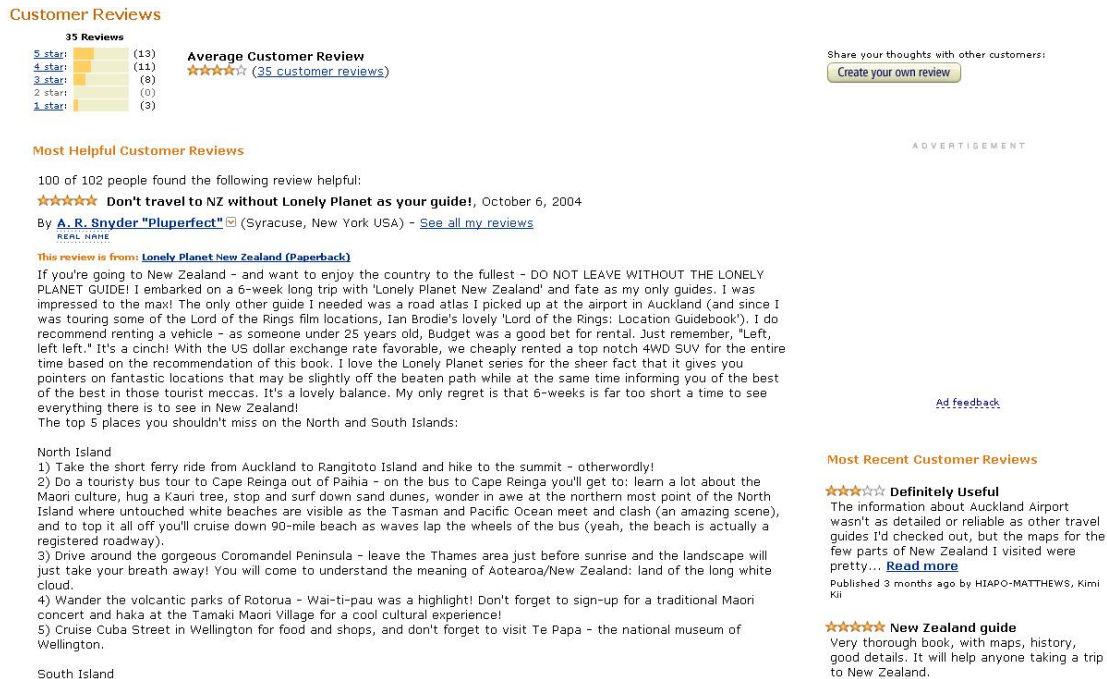


Figure 4.1c: An example of screenshot when customer clicks on a book item

The survey was administered to 128 respondents, consisting of 61 male and 67 female. All of the respondents have read online reviews in Amazon.com. In the survey, participants were given a scenario that they have to buy a travel guide for their trip to New Zealand. Based on the search results returned, they had to rank the order of information which they would search for so as to decide which book to click on and evaluate further. The survey results show that apart from price, ratings would be the first thing they would search for, followed by the text comments (see Figure 4.1d). Therefore,

for majority of the participants, they would first look at the ratings to decide which book to click on and evaluate further based on the customer reviews. Interestingly, 78.1% of the respondents agree that ratings are important in the filtering of their search results while sentiments are important in making their final choice. We further conducted a second survey with 156 respondents (100 male 56 female). A series of questions were asked as to the relative importance of numerical ratings and text reviews during the search, evaluation and purchase. From Figure 4.1e and 4.1f, the survey results on the importance of rating and sentiments for each stage of the decision making process reaffirmed the responses we obtained from our interview.

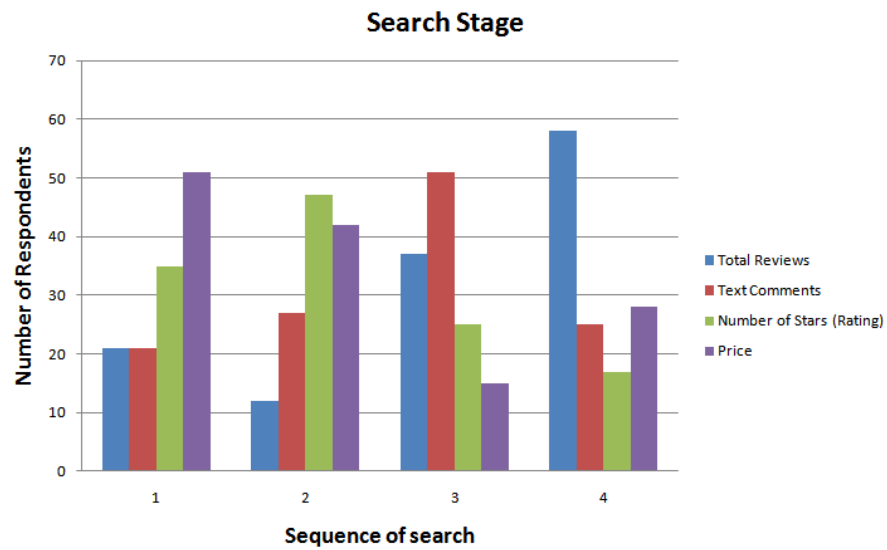


Figure 4.1d: Sequence of search for information

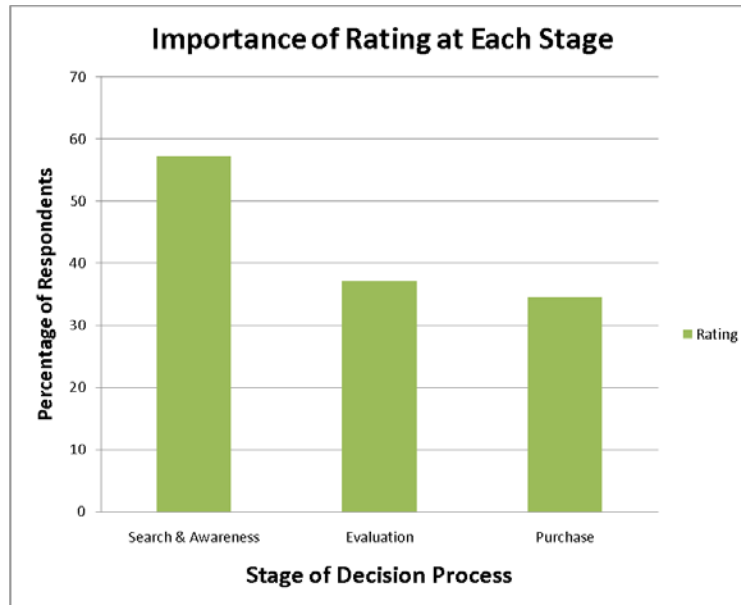


Figure 4.1e: Survey results on the relevance of rating

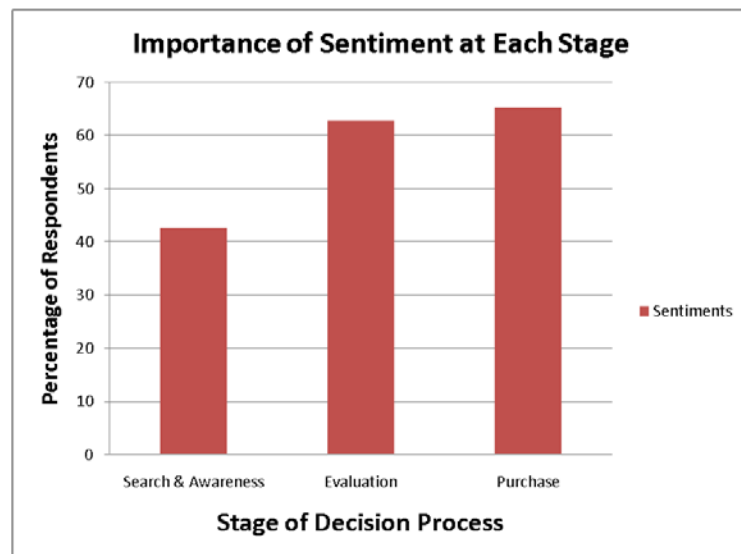


Figure 4.1f: Survey results on the relevance of sentiments

4.2. Online Reviews Format: Impact of Sentiments on Sales

In light of the above survey results, we proceed with a comprehensive study to empirically validate if sentiments do in fact have a greater influence on the purchase. In this section, we develop models to understand the impact online reviews have on

product sales. Although it would be ideal to have actual sales data on the books in our sample, only sales rank data is available in Amazon. Based on Schnapp and Allwine's (2001) finding, the relationship between log sales and log sales rank is linear. Other researchers have also found it to be useful to use log sales rank in their models (Chevalier and Mayzlin 2006, Forman et al. 2008, Li and Hitt 2008). Hence, we also chose log of sales rank as a proxy for product sales in our log-linear regression models.

The dependent variable is $\log(\text{Salesrank})$, which is the log of sales rank of product j in time t . μ_{jt} is a product fixed effect that controls for unobserved heterogeneity across 60% the book items (the entire dataset would have been too large for fixed effect analysis resulting in out of memory errors, therefore a subset is extracted. Nevertheless, our results are qualitatively similar if we used the entire dataset without fixed effects). The control variables used in our model include the Amazon retail price ($Price$), the difference between the date of data collection and the release date of the product (Age) and the log of the number of reviews posted for that product ($TotalReviews$). This is consistent with prior work such as Chevalier and Mayzlin (2006) and Forman et al (2008). In addition, the work of Forman et al. (2008) and Ghose and Ipeirotis (2010) indicates that consumer sales are also affected by the disclosure of the identity of the reviewer, hence, we include the average helpfulness ($AvgHelpful$) where $Helpful_{ji}$ is the ratio of helpful votes to total votes received for review i for product j . To account for potential non-linearities and to smooth large values, we take the log of sales rank and some of the control variables such as price, volume of reviews and age so as to be consistent with the literature (Forman et al. 2008, Ghose and Ipeirotis 2010).

Next, we estimate the effect of numerical ratings and its interaction with sentiment score on sales. The descriptive statistics and description of measures used in our study are presented in Table 4.2a and 4.2b.

Table 4.2a: Descriptive Statistics of Books

Variable	Median	Mean (SD)
Retail Price	16.47	29.39 (30.59)
Sales Rank	122,103	204,331 (243,085)
Age of Product (days)	1,685	2,414 (2,553)
Average Rating	4.5	4.35 (0.73)
Number of Reviews	4	15.74 (76.10)
Average helpful ratio	0.83	0.79 (0.20)
Average title score	4	3.97 (0.82)
Average content score	4.16	4.13 (0.51)
Average sentiment score	4.08	4.05 (0.55)
Average strong positive score	0.14	0.16 (0.12)
Average strong negative score	0	0.02 (0.03)
Average ordinary positive score	0.6	0.60 (0.15)
Average ordinary negative score	0.21	0.21 (0.12)

Table 4.2b: Description of measures used in the study

Measures	Description
<i>SalesRank</i>	<i>Sales rank on Amazon.com for book j</i>
<i>Price</i>	<i>The price on Amazon.com for book j</i>
<i>Age</i>	<i>The difference between the date of data collection and the release date of book j</i>
<i>TotalReviews</i>	<i>The number of reviews posted for book j</i>
<i>AvgHelpful</i>	<i>The average helpfulness for book j</i>
<i>AvgSentiScore</i>	<i>The average sentiment for book j</i>
<i>AvgTitleScore</i>	<i>The average title sentiment for book j</i>
<i>AvgConScore</i>	<i>The average content sentiment for book j</i>
<i>Avg Rating*Senti</i>	<i>The interaction between average rating and sentiment for book j</i>
<i>AvgStrPosScore</i>	<i>The average strong positive sentiment for book j</i>
<i>AvgStrNegScore</i>	<i>The average strong negative sentiment for book j</i>
<i>AvgOrdPosScore</i>	<i>The average ordinary positive sentiment for book j</i>
<i>AvgOrdNegScore</i>	<i>The average ordinary negative sentiment for book j</i>

To start off, we have a base model Model 1 that examines the impact of retail price, total number of reviews, age of the product, average helpfulness of reviews and average sentiment score on sales rank. We estimate product-level fixed effects to control for differences across products. These fixed effects are equivalent to including a dummy for every product in our sample, and so this allows us to control for differences in the average quality of products (Ghose and Ipeirotis 2010).

Prior work has shown that review valence (i.e. average rating) may be correlated with product-level unobservable that may be correlated with sales. In our setting, although we have controlled for differences in the average quality of products through fixed effects, it is possible that changes in the popularity of the product over time may be correlated with changes in review valence. Thus, this parameter may also reflect exogenous shocks that may influence product popularity (Forman et al. 2008, Ghose and Ipeirotis 2010). The variable on the volume of reviews (TotalReviews) will also capture changes in product popularity or perceived product quality over time; therefore, β_3 (in Model 1) may reflect the combined effects of a causal relationship between number of reviews and sales (Duan et al. 2008) and changes in unobserved book popularity over time (Ghose and Ipeirotis 2010).

Model 1

$$\ln(\text{SalesRank}_{jt}) = \beta_1 \ln(\text{Price}_{jt}) + \beta_2 \ln(\text{Age}_{jt}) + \beta_3 \ln(\text{TotalReviews}_{jt}) + \beta_4 (\text{AvgHelpful}_{jt}) + \beta_5 (\text{AvgSentiScore}_{jt}) + \mu_{jt} + \varepsilon_{jt}$$

Thereafter, we take an incremental approach by adding new variables into Model 1 to study their respective incremental effect. In Model 3, we add the average rating variable to the base model to study the incremental contribution of the rating variable. In Model 4,

we add the interaction term between the numeric rating and sentiments to study the synergistic impact both numerical ratings and sentiments have on product sales.

Finally, we study the impact of sentiments in detail. We decompose the sentiments into title and content. Then, we obtained the average title score and average content score for all the items and include them in Model 5.

In Table 4.2c, the results of Model 1 show that the price coefficient is positive and significant indicating when prices rise the sales rank increases (sales decline), an intuitive and a consistent finding. Chevalier and Mayzlin (2006) and others (Forman et al. 2008, Li and Hitt, 2008) find a similar result in their models of Amazon data on books. The coefficient associated with age is also positive and significant. This suggests that as the book remains longer on Amazon.com, its sales rank will be larger and sales will fall.

Table 4.2c: Model Comparisons (Dependent Variable: $\ln(\text{SalesRank})$) with fixed effects (standardized coefficients)

	Model 1	Model 2	Model 3	Model 4	Model 5
ln(Price)	0.1175***	0.1025***	0.1116***	0.1121***	0.1148***
ln(Age)	0.2838***	0.2793***	0.2846***	0.2844***	0.2860***
ln (TotalReviews)	-0.4854***	-0.4799***	-0.4877***	-0.4874***	-0.4917***
AvgHelpful	-0.0334***	-0.0328***	-0.0059***	-0.0037***	-0.0043***
AvgSentiScore	-0.2245***		-0.1943***	-0.1233***	
AvgRating		-0.1151***	-0.0671***	-0.0155*	-0.0600***
AvgRating*Senti				-0.0159***	
AvgTitleScore					-0.0433***
AvgConScore					-0.2157***
Intercept	10.6623***	10.3238***	10.8552***	10.6286***	11.0862***
Adjusted R²	0.2102	0.2072	0.2111	0.2112	0.2126
N	30222	30222	30222	30222	30222

Legend: *** $p < .001$; ** $p < .01$; * $p < .05$; + $p < .10$

The coefficient associated with the total number reviews is negative and significant implying that the higher number of reviews for an item the lesser the sales rank and greater the sales. This coefficient by far is the largest in all our models indicating the substantial influence that increasing number of reviews play in impacting sales. The dynamic flow of online reviews and their impact on product sales over time could be an interesting area of research for the future.

One other variable that we used in our models as a covariate is the average helpfulness of a review. Just above each review, Amazon.com provides a summary of the number of people who have read the review and found it helpful. We have computed the average usefulness of all reviews for an item. The coefficient is negative and significant indicating that as the average usefulness of reviews for item increases, sales rank declines (sales increase). One should note however, that the magnitude and the significance levels of this variable drop when numerical rating variable and sentiment rating scores are included in the model.

The sentiments expressed in online reviews have a significant negative impact on product sales rank. This suggests the positive influence of sentiments on product sales. Model 1 with the covariates and just the sentiment scores captures 21.02 % of the variance. When the average numerical ratings variable is added to the model (Model 3), two things happen. One, the ratings coefficient is negative and significant (-0.0671) just like the sentiments. Second, the coefficient associated with sentiments drops slightly from -0.2245 to -0.1943. This suggests that as one would expect that the two variables associated with online reviews, namely the summarized piece of evaluative information on the item as reflected in the numerical ratings and the

sentiment score extracted from the text of the review, both are influential in impacting product sales and they possess a degree of shared variance. The size of the coefficients provides us an indication of the relative impact of these two variables. In Model 3 when both the variables are included, the impact of sentiments is more than twice that of than ratings. Evidence of the joint influence is even more obvious when we introduce the interaction term of ratings and sentiments (Model 4). The coefficient associated with the interaction term is significant and negative. In this model, the ratings variable coefficient becomes less significant. The sentiment score coefficient continues to be negative and significant but its value has dropped to -0.1233 (from -0.1943 in Model 3). These results suggest that there is interaction effect between ratings and sentiments but this has not been reported in the previous literature.

Many times, the title of the review presents a summary view of what is in the full text of the review. Customers look at the titles of the reviews to get a feel for what the review might say and then decide to take the decision to read the text of the review. To see if there is a differential impact of the sentiments in the title and the content of the review, we decomposed the total sentiment score into that based on just the title of the online review (AvgTitleScore) and the one based on the entire review (AvgConScore). Using these two variables instead of the sentiment score improves the explained variance to 21.26%. Both variables show significant negative impact on sales rank suggesting that customers are influenced by the sentiments in the title as well as sentiments in the content of the online review. Although, the title sentiment score is significant, the coefficient associated with the content sentiment score is about 5 times that of title sentiment score indicating a substantially bigger sales

impact. Although title may convey some information that is useful, customers seem to be paying more attention to the sentiments expressed in the content of the review.

Table 4.2d: Model Comparisons (Dependent Variable: $\ln(\text{SalesRank})$) 5% Random Sample (standardized coefficients with fixed effects)

	Model 1	Model 2	Model 3	Model 4	Model 5
$\ln(\text{Price})$	0.1494***	0.1325***	0.1449***	0.1451***	0.1505***
$\ln(\text{Age})$	0.3081***	0.3062***	0.3086***	0.3086***	0.3103***
$\ln(\text{TotalReviews})$	-0.5058***	-0.5031***	-0.5077***	-0.5076***	-0.5109***
AvgHelpful	-0.0562*	-0.0776*	-0.0334*	-0.0330*	-0.0311*
AvgSentiScore	-0.2252***		-0.1987***	-0.1853*	
AvgRating		-0.1121***	-0.0629**	-0.0531*	-0.0555*
AvgRating*Senti				-0.0030*	
AvgTitleScore					-0.0335**
AvgConScore					-0.2402**
Intercept	10.4192***	10.0536***	10.6058***	10.5614***	10.8689***
Adjusted R²	0.2309	0.2275	0.2316	0.2316	0.2337
N	2515	2515	2515	2515	2515

Legend: *** $p < .001$; ** $p < .01$; * $p < .05$; + $p < .10$

The models and the results described so far are based on our large dataset of over 30,000 books. One might question that the significance levels seen here are artifacts of these large sample sizes. To address this issue and see how a smaller sample of books that has been typically analyzed previously will hold up our results, we have randomly sampled 5% of the total data giving us a sample of 2515 books. This compares with Chevalier and Mayzlin (2006) sample of 2387, Forman et al. (2008) sample of 3139 and Li and Hitt (2008) sample of 2203. We estimated all the five models and the results are presented in Table 4.2d. The results are similar to those discussed and presented in Table 4.2c.¹⁸ This presents us with an assurance of the robustness of our findings.

¹⁸ We have also conducted differences-in-differences analysis to control for unobservable shocks and the findings are consistent.

4.3. Robustness Check

4.3.1. On a different sentiment mining technique

In a previous section, when we discussed how we computed sentiment scores, sentiment terms are assigned to be strong or weak and given a higher weight for the strong sentiments as a heuristic measure. In this section, we draw upon the work of Archak et al. (2007) to check the robustness of our technique. For this part of the study, instead of grouping the sentiment terms into either strong or ordinary types, we estimate a weight for each individual sentiment term. Similarly, the list of words from the dictionary and those manually extracted will form our original list of sentiment terms. Based on this list of seed words, we proceed to determine the weight for each of the sentiment term. (Please refer to Chapter 3, Section 3.1.1.)

In this robust check, we estimate a different set of weights for the sentiment terms. First, we extract product items that have sales rank from 1 to 100,000. This is because sales rank within this range is updated daily and are therefore more accurate than items with sales rank above 100,000.¹⁹ Then from these extracted items, 30% of the dataset is used to perform the training (training set), the other 30% for tuning the parameters (tuning set) and the remaining of the dataset for testing (test set). For the training set, we calculate the frequency of sentiment term occurrences for each review. The next step is a summation of these term frequencies to the product item level. The final term-review matrix is normalized to a scale of 0 to 1. Then, we estimate the weight of each sentiment

¹⁹ We have also tried the training and tuning procedure by randomly selecting items from the whole dataset without the sales rank classification. The results obtained are qualitatively similar to those in Table 7, 8 and 10. In sum, we the results show that the sentiment has a much greater correlation on sales rank than numeric ratings; the sentiments in the content are much more impactful than those in the title; and ordinary sentiments are much more impactful than strong sentiments.

term using on the following regression model by controlling the price, total reviews and average rating for each product:

$$\ln(\text{SalesRank}) = \alpha_1 \ln(\text{Price}) + \alpha_2 \ln(\text{TotalReviews}) + \alpha_3 (\text{AvgRating}) + \alpha_4 \text{sentiment_term}_1 + \dots + \alpha_{n+4} \text{sentiment_term}_n + \varepsilon$$

The same procedure is carried out for the tuning set. Finally, the parameter estimate of each sentiment term from the training set and the tuning set are averaged to derive the final weight for each sentiment term. Sentiment terms with inconsistent polarity i.e. negative or positive are dropped. For each review, the sentiment score is calculated by the weighted sum of the number of positive terms minus the number of negative terms:

$$\frac{\sum_{i=1}^n \text{positive_term}_i * \text{wg}_i - \sum_{i=1}^m \text{negative_term}_i * \text{wg}_i}{\sum_{i=1}^n \text{positive_term}_i * \text{wg}_i + \sum_{i=1}^m \text{negative_term}_i * \text{wg}_i}$$

where:

- positive_term: the number of positive terms
- negative_term: the number of negative terms
- wg: the weight of each sentiment term

In the estimation of individual sentiment term weights, we present some of the results in Table 4.3.1a. These observations are consistent with those obtained by Archak et al. (2007). Table 4.3.1b shows the results of the same 5 models discussed in the earlier section, but using the sentiment data using the optimal weighting scheme described earlier. As is clear from these results, even in the case when each sentiment term has its own individual weight, the findings obtained are qualitatively similar to the first technique when an arbitrary weight is given for the strong and

ordinary sentiment terms. On the whole, our results show three key highlights: 1) the variance explained by sentiments is much higher than that of ratings; 2) the sentiments in the content impact sales more compared to the sentiments in the title alone and 3) the interaction between ratings and sentiments is statistically significant on sales rank. Hence, the findings in Table 4.3.1b are consistent with that of the previous section which uses the heuristic measure.

Table 4.3.1a: Some sentiment term weight

Sentiment term	Weight
excellent	-0.26544
fantastic	-0.48469
super	-2.33820*
brilliant	-0.88951*
interest	-0.69542*
pretty	-0.75699 ⁺
nice	-0.05906
useless	0.91457*
boring	0.40625
pleasure	0.91201
fabulous	0.68397
absurd	2.51077*
grievance	10.01445 ⁺

Legend: *** $p < .001$; ** $p < .01$; * $p < .05$; ⁺ $p < .10$

Thus, we have demonstrated that a simple heuristic weighting method works just as well as a weight calibration process. As the amount of user-generated content will be voluminous, the heuristic process will help code and use our large-scale data without resorting to sampling a small portion of it for analysis.

Table 4.3.1b: Model Comparisons (Dependent Variable: $\ln(\text{SalesRank})$)
(standardized coefficients)

	Model 1	Model 2	Model 3	Model 4	Model 5
ln(Price)	0.0586***	0.0561***	0.0553***	0.0554***	0.0557***
ln(Age)	0.1486***	0.1498***	0.1500***	0.1500***	0.1502***
ln (TotalReviews)	-0.3204***	-0.3201***	-0.3254***	-0.3253***	-0.3256***
AvgHelpful	-0.0096 ⁺	-0.0099 ⁺	-0.0014 ⁺	-0.0012 ⁺	-0.0012 ⁺
AvgSentiScore	-0.0664***		-0.0510***	-0.0322**	
AvgRating		-0.0553***	-0.0442***	-0.0273*	-0.0449***
Avg Rating*Senti				-0.2810*	
AvgTitleScore					-0.0136***
AvgConScore					-0.0485***
Intercept	0.2649***	0.2648***	0.2645***	0.2646***	0.2642***
Adjusted R²	.2354	.2354	.2354	.2356	.2360
N	20120	20120	20120	20120	20120

Legend: *** $p < .001$; ** $p < .01$; * $p < .05$; ⁺ $p < .10$

4.3.2. On a different dataset

Sales Rank to Sales Quantity Calibration

To verify the robustness of the results, we examine another two extensions. The first extension verifies that the transformation of the sales rank to actual sales quantity (based on prior works' estimated parameters) will not change the qualitative nature of the results. The second extension went beyond our existing cross-sectional dataset. Using the actual point-of-sales data provided by Neilson Bookscan, we check the veracity of our results.

Historical work in this area such as that of Chevalier and Mayzlin (2006), Archak et al (2007), Forman et al. (2008) have used sales rank as the dependent variable in their analyses. Such analyses is possible because prior research in marketing such as Chevalier and Goolsbee (2003), Ghose, Ipeiritos and Sundararajan (2007) has associated sales ranks with demand levels for products in Amazon. The association is

based on the experimentally observed fact that the distribution of demand in terms of sales rank has a Pareto distribution (i.e. a power law) (Chevalier and Goolsbee). Based on this observation, it is possible to convert sales ranks into sales quantity levels using the following Pareto relationship:

$$\ln Quantity = \alpha + \beta \ln Rank$$

Where Q is the unobserved sales quantity, Rank is its observed sales rank and $\alpha > 0$, $\beta < 0$ are industry-specific parameters. Therefore, the log of product sales rank on Amazon.com can serve as a proxy of the log of product demand.

Beyond our current dataset, we also obtained the Neilson BookScan actual point-of-sales data for the Top 1000 books that correspond to the similar time period as our cross-sectional dataset. Using the log of sales quantity provided by Neilson Bookscan as the dependent variable, we mapped each book item's sales with the Amazon's sales rank to obtain the α and β parameter. From our dataset, the estimated α parameter is 10.2589 while β parameter is -0.1705. Again, we see qualitatively similar results obtained from the Neilson Bookscan point-of-sales dataset in Table 4.3.2. Varying the parameter estimates has no impact on the directional results as it is a linear transformation. On the whole, our results still hold and the conclusion remains the same:

1. the variance explained by sentiments is much higher than that of ratings;
2. the sentiments in the content impact sales more compared to the sentiments in the title alone;
3. the interaction between ratings and sentiments is statistically significant on sales.

Table 4.3.2: Neilson Bookscan Sales Data with Fixed Effects
Dependent Variable = Log Sales Quantity

	Model 1	Model 2	Model 3	Model 4	Model 5
ln(Price)	-0.1587**	-0.1697**	-0.1583**	-0.1393**	-0.1614**
ln(Age)	-0.0689**	-0.0562**	-0.0687**	-0.0669**	-0.0667**
ln (TotalReviews)	0.0709***	0.0624***	0.0706***	0.0797***	0.0701***
AvgHelpful	0.2744**	0.2018**	0.2726**	0.1869**	0.2681**
AvgSentiScore	0.2556**		0.2572**	0.7265***	
AvgRating		0.0278**	0.0050*	0.2042**	0.0043**
AvgRating*Senti				0.0910**	
AvgTitleScore					0.0755**
AvgConScore					0.1635**
Intercept	7.7755***	8.6086***	7.7930***	6.6449***	7.8527***
Adjusted R²	0.05	0.04	0.06	0.06	0.06
N	843	843	843	843	843

Legend: *** $p < .001$; ** $p < .01$; * $p < .05$; + $p < .10$

4.4. Implications

In this study, we consider the distinctive online review design in which sentiments are expressed. We demonstrate that sentiments expressed in the text of online reviews have a significant impact on product sales. We were able to tease out the partial effects of ratings and sentiments on product sales and we found that sentiments explained greater variance and had a substantially stronger effect even in the presence of numerical ratings in the model. The significant interaction between sentiments and numerical ratings show that customers use both of them in arriving at the final choice.

We also showed that there is a differential effect of sentiments expressed in the title of the review and the content of the review. The impact of content sentiments is twice as large as the title sentiments. It is likely that in this case, customers are using title sentiments as a screening device but still would validate their choice by digging into the content sentiments. One possible implication for reviewers here is that they may need to pay attention to the way the title of the review is written. It should be

crisp, and clearly pointing to the sentiments expressed in the full text so that it makes it attractive for potential buyers to look deeper at their review.

Finally, based on in-depth interviews with online shoppers and our experiential survey results, 78.1% of the participants agree that ratings are important in their search for information and in filtering of the search results. Our empirical findings support our survey results and have shown the stronger influence of sentiments over ratings on the evaluation and final purchase. This study shed insight on the *relevance* of rating and sentiments over different stages of the consumer decision making process and advances the understanding of the consumer decision making process by looking at how ratings and sentiments affect consumers' purchase decision.

Using a large scale cross-sectional data, and subsequent robustness check using different data sources and sentiment mining technique, the rigor of these analyses present consistent results, which are compelling and directional to managers and marketers.

Chapter 5.

Manipulation of Online Reviews

As consumers become increasingly reliant on online reviews to make purchase decision, the sales of the product becomes dependant on the word of mouth that it generates. As a result, there can be attempts by firms to manipulate online reviews to increase their product sales. Despite the existence of such activity, the amount of such manipulation is unknown, and deciding which reviews to believe in is largely based on the reader's discretion and intuition. In this study, we propose a simple statistical method to detect online reviews manipulation and assess how consumers respond to such manipulation. In particular, the writing style of reviewers is examined and the effectiveness of manipulating through ratings, sentiments and readability is investigated. We discover that about 10.3% of the products are subjected to online reviews manipulation. In spite of the deliberate use of manipulative sentiments and ratings in fraud reviews, consumers are only able to

detect manipulation activity through ratings, but not through sentiments. The findings from this research ensues a note of caution for all consumers that make use of online reviews of books for making purchases, and encourage them to delve deep into the book reviews without getting trapped in the fraudulent manipulation.

In the next section, the methods of detecting manipulation will be discussed and examples of such activity will be shown. To check if our method works in detecting manipulation, the evidence of manipulation discovered by our technique is presented in Section 5.2. Subsequently, based on the items discovered to be manipulated, we analyze the impact on sales in section 5.3. Finally, Section 5.4 concludes with the implications.

5.1. Manipulation detection

Consumers are increasingly relying on opinions posted in online reviews to make a variety of decisions ranging from what movies to watch to what stocks to invest in (Guernsey 2000). Previously, these decisions were based on advertisements or product information provided by vendors. However, with the proliferation of e-commerce and increasing number of product reviews provided by users, it has been found that consumers have switched to online reviews for their search on information related to a variety of products. Prior research has also found that consumers find such user-generated reviews more credible and trustworthy than the traditional sources (Bickart and Schindler 2001). However, it is generally not known to what extent these online reviews are truthful ‘user-generated’ reviews or merely reviews provided by vendors interested to push the sales of products. In addition, it is not clear how effective are the manipulation of online reviews in influencing consumers’ purchase

decisions. Therefore, the second study of this dissertation is to develop a method to identify manipulation activity and analyze the influence of this action.

5.1.1. Existence of manipulation activity

Manipulation of reviews occurs when online vendors, publishers, or authors write “consumer” reviews by posing as real customers. Thus, manipulation here means that the review posted is not a truthful account of a real customer’s experience. Manipulation of reviews is not a hypothetical phenomenon. It is known to exist widely in popular websites related to e-commerce, travel, and music. For example, when Amazon.com’s Canadian website accidentally revealed the true identities of some of its book reviewers due to software errors, it was found that a sizable proportion of these reviews were written by the book’s own publishers, authors and their friends or relatives (Harmon 2004). Figure 5.1.1 shows a case in which one reviewer plagiarized the content of another product for his or her products.

...In this well-researched, entertaining, and immensely readable book, Pinch (science & technology, Cornell Univ.) and Trocco (Lesley Univ., U.K.) chronicle the synthesizer's early heady years, from the mid-1960s through the mid-1970s.....Throughout, their prose is engagingly anecdotal and accessible, and readers are never asked to wade through

...In this well-researched, entertaining, and immensely readable book, Kettlewell chronicles the synthesizer's early, years, from the turn of the 20th century - through the mid-1990s.....Throughout, his prose is engagingly anecdotal and accessible, and readers are never asked to wade through dense, technological jargon. Yet there are enough details to enlighten those trying to understand this multidisciplinary field of music, acoustics, physics, and electronics. Highly recommended³

Figure 5.1.1: An Example of Manipulation of Online Reviews

Review manipulation is not prevalent just amongst book sellers. The music industry is known to hire professional marketers who surf various online chat rooms and fan sites to post positive comments about new albums (Mayzlin 2006; White 1999). Insiders of the travel industry claimed that reviews in their industry have been manipulated; either by the owners or competitors.²⁰ A former restaurateur revealed how he leveraged TripAdvisor to increase his business. “I just wanted to give you my input on my experience as a business owner who artificially ‘upped’ my own rating...I began tracking feedback about my restaurant on TripAdvisor “rants and raves” page. It very quickly occurred to me that I could [write] in glowing reviews about my own restaurant and up my ratings numbers... After a period of time, I began to see my rating slide a bit after some not so positive postings by supposedly “real” customers. The complaints that were written about seemed somewhat contrived.....Were they posted by my competition? Perhaps, but I didn’t let it concern me too much. I simply got on TripAdvisor and bombarded them with glowing reviews about my own restaurant! Within days, I was rated a perfect 5!” Evidently, manipulative activity also exists widely in the tourism industry. The well-known publisher of travel guides Frommers remarked: “Why wouldn’t a hotel submit a flurry of positive comments penned by employees or friends? If you were a hotel owner, wouldn’t you take steps to make sure that TripAdvisor contained numerous favorable write-ups of your property?”²¹

²⁰ <http://www.tripso.com/today/new-tripadvisor-whistleblower-claims-some-reviews-are-totally-fraudulent/>

²¹ <http://www.elliott.org/blog/does-tripadvisor-hotel-manipulation-scandal-render-the-site-completely-useless/>

The various pieces of discussion in the above paragraph have shown that online review manipulation is a common industry practice and a serious problem. The consequence of this is consumers may make the wrong purchase decision based on these manipulated information and firms who made the wrong decisions based on fraud reviews posted by their competitors. Yet, there have been few empirical studies that have investigated and reported the presence of manipulated reviews in the online review forums. To our best knowledge, there are only two recent empirical works focusing on proving the existence of online review manipulation without offering ways to identify manipulation in reviews (Hu et al. 2010a and Hu et al. 2010b). The development of a technique to identify manipulation is a research challenge.

Since participants of online review communities can assume any identities or choose to remain anonymous, marketers are able to disguise their promotion of products as consumer recommendations. In an online context, if potential customers knew which reviews were posted by real customers who consumed the product, and which reviews were written by authors, publishers, or any third parties with selfish interests, then those potential customers could undo the damages caused by these slanted reviews and dismiss such reviews. Unfortunately, since most intended manipulative reviews were written by anonymous entities or by manipulators who assumed a customer's identity, it is difficult to differentiate an enthusiastic review from a manipulated review. Even a manual inspection of the content of a review is still difficult to differentiate between truthful and manipulated reviews unless some parts of the manipulated review were identical to another review (David and Pinch 2005).

In this study, we set off to discover the presence of manipulation in online reviews of products and identify the effectiveness of such manipulative activity on the sales of products.

We specifically address the following research questions:

- 1. What is the extent of manipulation that is present in online reviews?*
- 2. How can such manipulation be detected from the ratings and textual content of reviews? What are some of the textual characteristics that can be used to identify products with manipulated reviews?*
- 3. What is the impact of review manipulation in terms of rating and writing style on the sales of products?*

To answer the above questions, we first describe the intuition behind the method for the detection of promotional reviews. As writing style varies with the background of an individual, intuitively, reviews written by different consumers will be random in the case where there is no manipulation (Holmes 1994, Hu et. al 2010).

In our context, writing style refers to how consumers construct sentences together when they write online reviews. Reviews written by individual consumers often express a personal view of their experience on the products. Thus their writing style should be different from one another. And such differences reflect the heterogeneity among their culture, education, occupation and so on. However, for manipulators, the situation is different. If reviews are consistently monitored and posted by manipulators, then the observed reviews will be a blend of true customer reviews and manipulators' reviews; hence the writing styles of observed reviews will not be random with the existence of manipulators.

By observing the change in the writing style over time, we can infer whether the online reviews for a product is manipulated or not because writing style is unique among individuals. Building on this intuition, we develop a model for the detection of manipulation.

5.1.2. Randomness of writing style

A consumer review consists of two parts: a numerical rating of the product or service being reviewed, as well as textual statements on the product or service. The intuition is that when unethical users manipulate online reviews, they either post reviews with a high numeric rating or manipulate the textual statements posted in the review. If reviews are indeed written by different customers, the writing style should be random. Hence, by investigating if there is randomness in the rating or writing styles of reviews posted over time, we may be able to detect manipulation in online reviews.

We focus on two different ways of evaluating writing styles - *Sentiments and Readability*. In the attempt to write reviews that customers will believe and act upon, manipulators are likely to use certain persuasion strategies. Persuasion is the use of appeals to convince a listener or reader to think or act in a particular way. In ancient Greece, the art of using language as a means to persuade was called rhetoric. The Greek philosopher Aristotle (384-322BC) set forth an extended treatise on rhetoric that still attracts great interest and careful study even today. His treatise on rhetoric discussed not only the elements of style and delivery, but also emotional appeals (pathos) and character appeals (ethos) (Garsten 2005). He identified three main forms of rhetoric:

- ethos: how the character and credibility of a speaker/writer could influence an audience to consider him/her to be believable.
- pathos: the use of emotional appeals to alter the audience's judgment. This could be done through the use of metaphors, emotive language and sentiments that evoked strong emotions in the audience.
- logos: the use of reasoning to construct and support an argument (e.g. use of statistics, mathematics, and logic).

Since ethos is not applicable in our context as posters of reviews are mostly anonymous, our focus here will be on pathos i.e. the use of emotive language. In regard of logos, the application of text mining techniques to test the logic and reasoning of reviews would be a challenging and interesting future work.

In the online review environment, manipulators are likely to use sentiments to slant reviews (i.e., write or present in a biased manner) so as to influence a potential reader's purchase behavior. The use of such a slanting behavior is common in public relations, lobbying, law, marketing, professional writing and advertising where the goal of the writer is to influence the third party's opinion or belief. For example, Kahn and Kenney (2002) conducted content analysis of campaign coverage in major newspapers for 67 incumbent Senate campaigns between 1988 and 1992, and found that the papers' editorial endorsements significantly affected the tone (i.e., positive, neutral, negative) of the incumbent coverage and the number of criticisms published about incumbents; and such editorial slants in turn influenced voters' decisions in the elections. Likewise, Gurun and Butler (2009) found that when local media reported news about local companies, they used fewer negative words than when they reported

about non-local companies. As the local companies spent more on advertising, the local media had more positive slant towards them. The researchers reported that on an average, an increase in local media slant by one standard deviation was associated with a 3.59% increase in the market value of the firm. From these examples it might be reasonable to think that in the context of online reviews, manipulators would tend to use positive slant in the form of positive sentiments to persuade and influence customers' choices.

In addition to the sentiments of writing style, another important metric that will be used to discover manipulation is readability. Readability is defined as the ease of reading which will improve the comprehension as well as the retention of the textual material. Readability of textual data indicated the amount of effort that was needed by a person of a certain age and education level to understand a piece of text (Zakaluk and Samuels 1988). Readability is a score generated by a readability formula, and is derived from a mathematical model that assessed the reading ease of different pieces of text by a number of subjects. Based on the syntactical elements and the underlying style, the readability test would provide an indication of the understandability of a piece of text. In this study, we use the Automated Readability Index (ARI) (Senter 1967) and using the following formula:

$$ARI = 4.71 \left(\frac{\text{Total number of characters}}{\text{Total number of words}} \right) + 0.5 \left(\frac{\text{Total number of words}}{\text{Total number of sentences}} \right) - 21.43$$

The score obtained from readability tests represented the school grade level that was required to comprehend the piece of text and to understand the logic of the statement (for details please refer to Chapter 3 Methodology Section 3.3).

5.1.3. Wald-Wolfowitz (Runs) test

If reviews were indeed written by customers, then the writing style of the reviews would be random due to the diverse background of the customers. Therefore, a simple and intuitive way to detect the randomness of the review is to conduct a statistical test of randomness of writing styles and ratings of the reviews across time for each product that was reviewed. A non-random result in such a test would indicate the existence of manipulation. For this purpose, we adopted the Wald-Wolfowitz (runs) test to check the randomness of ratings, sentiments, and readability of the reviews over time.

The Wald-Wolfowitz test, also known as the Runs test for randomness, is used to test the hypothesis that a series of numbers is random (Gujarati 2003). The runs test is a non-parametric statistical test; therefore the interpretation of the results does not depend on any parameterized distributions. A “run” of a sequence simply refers to a segment consisting of adjacent equal elements. For example, the sequence:

++++-----++++-----+++++++-----

consists of 6 runs, three of which consist of + and the 3 on of -. To carry out the test, the total number of runs (R) is computed along with the number of positive and negative runs. To simplify the computations, the data are first centered on their mean.²² A positive run is determined as a sequence of values that are greater than zero, and a negative run is identified as a sequence of values that are less than zero. The number of positive runs (n) and negative runs (m) are checked to see if they are

²² We have also conducted the runs test for a non-normal distribution using median instead of mean as the reference point. The results have shown that the percentage of books with non-randomness within each sales rank category is qualitatively similar to that using the mean.

distributed equally in time. The test statistic is asymptotically normally distributed.

The large sample test statistic Z is given by:

$$Z = \frac{(R - E(R))}{\sqrt{V(R)}} \quad , \quad \text{where} \quad E(R) = \frac{2nm}{n+m} + 1 \quad , \quad \text{and} \quad V(R) = \frac{2nm(2nm - n - m)}{(n+m)^2(n+m-1)}$$

A finding of significance means that the series of reviews posted does differ significantly from random. The Runs test result for each product item is denoted on a binary scale of 1 and 0 where 1 represents non-random (with manipulation) and 0 represents random (without manipulation). Using the above procedure, we conducted the Runs test on ratings, sentiments, and readability of reviews for each product.

The final sentiment manipulation index $avg_senti_runs_i$ for any review i is computed as the average of the manipulation index of each type of sentiment term based on its Runs test using equation (5.1.3):

$$avg_senti_run_i = \frac{(str_pos_runs_i + str_neg_runs_i + ord_pos_runs_i + ord_neg_runs_i)}{4} \quad (5.1.3)$$

where $str_pos_runs_i$ is the runs test score for strong positive sentiments in review i , $str_neg_runs_i$ is the runs test score for strong negative sentiments in review i , $ord_pos_runs_i$ is the runs test score for ordinary positive sentiments in review i , and $ord_neg_runs_i$ is the runs test score for ordinary negative sentiments in review i .

5.2. Robustness Check

5.2.1. Evidence of manipulation discovered by Runs test

To verify if our Runs test method is able to detect manipulative activity, a manual inspection is conducted. Amongst all the items that were detected to have non-random reviews, we conduct a manual check to see if there are indeed reviews

posted by the *same person* for the *same book item*. From the items that were found to have non-random reviews, we have found an abundant evidence of such activities. Figure 5.2.1a to Figure 5.2.1d presents only a small number of the evidence found. ‘ASIN’ refers to the unique identification of a book while CustomerID is the unique identity of the customer. The figures show that there have been cases where an individual has posted several reviews for the same book item. These gave us confidence on the effectiveness of Runs test to detect manipulation of online reviews.

ASIN	AverageRating	TotalReviews	Rating	HelpfulVotes	CustomerId	TotalVotes	RevDate	
0385504209	3.5	3052	4	36	A1M4NJYPOVNL8Q	52	2004-05-08	Take Only As Dir
0385504209	3.5	3052	4	11	A16W9E27VW9IND	36	2004-05-05	Gripping and intri
0385504209	3.5	3052	5	27	A2MV5ADA3568DO	56	2004-05-04	A "Code" Worth I
0385504209	3.5	3052	3	37	A3TEH90X39WC8F	43	2004-04-30	What Makes a TI
0385504209	3.5	3052	5	13	A2SIE5S9T84J9	36	2004-04-29	I AM ENJOYING
0385504209	3.5	3052	1	35	A2JA9LYSXZES1A	83	2004-04-25	probably better if
0385504209	3.5	3052	1	35	A10ZH57J6QP844	46	2004-04-24	Buyer Beware
0385504209	3.5	3052	2	57	AU1XXY2S6FZQ2	86	2004-04-23	I dont get it
0385333218	5	491	5	0	A8LB47171JOQJ	0	2000-07-20	AN ADVENTURE
0385333218	5	491	5	0	ASNLJKAV3DBZX	0	2000-07-19	Funny, poignant,
0385333218	5	491	4	0	A1YVCJWWGCOIAI	0	2000-07-18	A Sweet Book
0385333218	5	491	5	0	ATVPDKIKX0DER	0	1999-10-04	One of the best b
0385333218	5	491	5	0	ATVPDKIKX0DER	0	1999-10-03	Just great
0385333218	5	491	5	0	ATVPDKIKX0DER	0	1999-10-03	The book was an
0385333218	5	491	5	0	ATVPDKIKX0DER	0	1999-10-02	A Wonderful Boo
0385333218	5	491	5	0	ATVPDKIKX0DER	0	1999-09-29	Dreams Beyond '
0385333218	5	491	5	0	AHVTCYHS5XSYM	0	1999-09-28	AN INSPIRATION
0385333218	5	491	5	0	ATVPDKIKX0DER	0	1999-09-25	THANKS FOR TH
0385333218	5	491	5	0	ATVPDKIKX0DER	0	1999-09-23	A wonderful book
0385333218	5	491	5	0	A3N8ITRDS67TVO	0	1999-09-23	Fantastic
0385333218	5	491	5	0	A3IX35WEY1Q21U	0	1999-09-09	I havent read a br
0385333218	5	491	5	0	ATVPDKIKX0DER	0	1999-09-03	THIS BOOK IS IE
0385333218	5	491	5	0	A2WVAQN7UM2LDW	0	1999-09-03	Wow!
0385333218	5	491	5	0	ATVPDKIKX0DER	0	1999-09-02	All childhood rocl
0385333218	5	491	5	0	A19EWP1U3X1I2B	0	1999-08-29	Inspirational stor
0385333218	5	491	5	0	A2THG37NXPJB6C	0	1999-08-28	A Superior Book
0385333218	5	491	5	0	ATVPDKIKX0DER	0	1999-08-28	Americana at its
0385333218	5	491	5	0	A2Z9YDOWG6VB0G	0	1999-08-28	If you read only o

Figure 5.2.1a: Evidence of Manipulated Reviews

ASIN	AverageRating	TotalReviews	Rating	HelpfulVotes	CustomerId	TotalVotes	RevDate	Summary
0060000074	5.0	55	5	2	A3UN6WX5RRO2AG	5	2004-10-10	Fantastically Fantastic
0060000074	5.0	55	5	2	A3UN6WX5RRO2AG	5	2004-10-10	The Best Book Ever!
0060000074	5.0	55	5	3	A3UN6WX5RRO2AG	5	2005-03-08	Fire saved our clan!
0060000074	5.0	55	5	2	A3UN6WX5RRO2AG	3	2005-01-18	The Darkest Hour
0060000074	5.0	55	5	2	A3UN6WX5RRO2AG	4	2004-10-26	Awesome.....
0060000074	5.0	55	5	5	A3UN6WX5RRO2AG	6	2005-02-06	The Darkest Hour
0060000074	5.0	55	5	2	A3UN6WX5RRO2AG	3	2004-11-25	An Amazing Story
0060000074	5.0	55	4	6	A3UN6WX5RRO2AG	11	2005-02-15	Loved the series
0060000074	5.0	55	5	3	A3UN6WX5RRO2AG	4	2004-12-12	Warrior 6: The Darkest Hour
0060000074	5.0	55	5	1	A3UN6WX5RRO2AG	8	2004-10-06	Number6
0060000074	5.0	55	5	3	A3UN6WX5RRO2AG	4	2004-10-20	Can Fire save save the clan?
0060000074	5.0	55	5	4	A3UN6WX5RRO2AG	5	2004-10-10	Terrific!!!
0060000074	5.0	55	5	3	A3UN6WX5RRO2AG	4	2005-04-24	A wonderful series, a wonderful book!
0060000074	5.0	55	5	6	A3UN6WX5RRO2AG	6	2005-03-24	A Stunning Read
0060000074	5.0	55	5	5	A3UN6WX5RRO2AG	6	2004-11-14	One of the best books EVER!!!!!!!!!!!!
0060000074	5.0	55	5	4	A3UN6WX5RRO2AG	4	2005-07-03	A Prophecy Completed
0060000074	5.0	55	5	1	A3UN6WX5RRO2AG	2	2004-11-07	Erin Hunter: Books sensational!!
0060000074	5.0	55	5	2	A3UN6WX5RRO2AG	8	2005-02-22	Great Book
0060000074	5.0	55	5	3	A3UN6WX5RRO2AG	4	2005-01-22	!!!!!!!!!!!!!!!!!!!!!!!!!!!!
0060000074	5.0	55	5	5	A3UN6WX5RRO2AG	7	2005-02-19	The Best Book Yet!!!

Figure 5.2.1b: Manipulated Reviews posted by the same customer for one book item

ASIN	AverageRating	TotalReviews	Rating	HelpfulVotes	CustomerId	TotalVotes	RevDate	Summary
0060739495	4.5	601	5	0	A3UN6WX5RRO2AG	2	2005-06-16	Never doubt a man until you walk two moons in his moccasins!
0060739495	4.5	601	5	3	A3UN6WX5RRO2AG	6	2005-04-13	Mrs. Browns 4th Grade Class
0060739495	4.5	601	5	1	A3UN6WX5RRO2AG	2	2005-03-16	so clever- everything fits together like a puzzle
0060739495	4.5	601	5	2	A3UN6WX5RRO2AG	5	2000-11-19	Walk Two Moons meant a lot to me!
0060739495	4.5	601	3	1	A3UN6WX5RRO2AG	2	2000-11-28	A review of walk two moons by a 6th grade reader
0060739495	4.5	601	4	2	A3UN6WX5RRO2AG	2	2000-11-29	Walk Two Moons
0060739495	4.5	601	5	5	A3UN6WX5RRO2AG	5	2000-12-07	The Best Book
0060739495	4.5	601	5	0	A3UN6WX5RRO2AG	1	2000-12-13	I love this book
0060739495	4.5	601	4	0	A3UN6WX5RRO2AG	1	2001-02-07	Great Book!!!
0060739495	4.5	601	1	0	A3UN6WX5RRO2AG	3	2001-03-20	Walk Two Moons: my opinion
0060739495	4.5	601	5	1	A3UN6WX5RRO2AG	9	2001-04-12	Beths Book Reveiw
0060739495	4.5	601	5	0	A3UN6WX5RRO2AG	4	2001-04-27	Travel the country, learn the truth
0060739495	4.5	601	4	1	A3UN6WX5RRO2AG	2	2001-05-12	Heres a Good Book!
0060739495	4.5	601	5	3	A3UN6WX5RRO2AG	3	2001-05-21	Walk Two Moons
0060739495	4.5	601	5	0	A3UN6WX5RRO2AG	1	2001-06-29	" DONT JUDGE A MAN UNTIL YOUVE TWO MOONS IN HIS MOCCASSINS"

Figure 5.2.1c: Evidence of Manipulated Reviews

ASIN	AverageRating	TotalReviews	Rating	HelpfulVotes	CustomerId	TotalVotes	RevDate	Summary
0064407314	4.5	466	4	0	A3UJ5K3T2T9BB6	1	2004-03-08	Monster
0064407314	4.5	466	5	1	A3UN6WX5RR02AG	1	2003-04-03	Monster
0064407314	4.5	466	5	0	A3UN6WX5RR02AG	1	2003-10-02	Great Book
0064407314	4.5	466	3	1	A3UN6WX5RR02AG	1	2003-11-12	MONSTER
0064407314	4.5	466	4	1	A3UN6WX5RR02AG	2	2004-05-27	My Book Review of Monster
0064407314	4.5	466	5	1	A3UN6WX5RR02AG	1	2004-11-10	Monster
0064407314	4.5	466	4	1	A3UN6WX5RR02AG	1	2005-01-26	Sad but Inspiring
0064407314	4.5	466	5	1	A3UN6WX5RR02AG	1	2005-02-17	The Best Book Ive Ever Read
0064407314	4.5	466	4	1	A3UN6WX5RR02AG	1	2005-03-25	Monster Review
0064407314	4.5	466	4	1	A3UN6WX5RR02AG	1	2005-03-25	Monsters in jail
0064407314	4.5	466	5	1	A3UN6WX5RR02AG	1	2005-03-25	A heart warming tale of a boy in prison
0064407314	4.5	466	4	1	A3UN6WX5RR02AG	1	2005-03-25	Its Straight
0064407314	4.5	466	4	1	A3UN6WX5RR02AG	1	2005-03-25	A typical monster
0064407314	4.5	466	4	1	A3UN6WX5RR02AG	1	2005-03-25	Complete and Total Admiration
0064407314	4.5	466	4	1	A3UN6WX5RR02AG	1	2005-03-25	Monster Review
0064407314	4.5	466	4	1	A3UN6WX5RR02AG	1	2005-03-25	Monster- by Z. Jaquandra Motisha
0064407314	4.5	466	5	1	A3UN6WX5RR02AG	1	2005-03-25	Jake Berenson
0064407314	4.5	466	4	1	A3UN6WX5RR02AG	1	2005-03-25	Bud country review
0064407314	4.5	466	5	1	A3UN6WX5RR02AG	1	2005-03-25	Monster Review
0064407314	4.5	466	5	2	A3UN6WX5RR02AG	2	2002-02-21	My review for Monster, by: Walter De Myers
0064407314	4.5	466	4	1	A3UN6WX5RR02AG	1	2002-03-20	MONSTER
0064407314	4.5	466	5	0	A3UN6WX5RR02AG	1	2002-06-04	Steve Harmon=Monster
0064407314	4.5	466	5	1	A3UN6WX5RR02AG	1	2002-11-25	Buy This Book NOW!
0064407314	4.5	466	5	0	A3UN6WX5RR02AG	1	2001-09-11	Monster,TLK
0064407314	4.5	466	5	1	A3UN6WX5RR02AG	2	2002-12-13	****Monster****
0064407314	4.5	466	5	5	A3UN6WX5RR02AG	6	2001-04-24	This Book Rocks.....Great Insite!!
0064407314	4.5	466	5	1	A3UN6WX5RR02AG	1	2001-06-03	Monster
0064407314	4.5	466	4	1	A3UN6WX5RR02AG	1	2001-06-04	Monster By Walter Dean Myers
0064407314	4.5	466	4	0	A3UN6WX5RR02AG	1	2001-10-30	Monster by David N.
0064407314	4.5	466	4	1	A3UN6WX5RR02AG	2	2001-11-30	Monster
0064407314	4.5	466	5	0	A3UN6WX5RR02AG	1	2003-01-16	An exciting book
0064407314	4.5	466	5	1	A3UN6WX5RR02AG	1	2002-12-13	MY REVIEW ON MONSTER

Figure 5.2.1d: Evidence of Manipulated Reviews

5.3. Impact of Manipulation on Sales

The data used in this research were gathered from Amazon.com. The reason for picking Amazon.com for the dataset was because past research had investigated manipulation of online reviews for this site (David and Pinch 2005) which is also supported by the evidence shown in our dataset in Section 5.2.1. To have a meaningful Runs test, we retained books that had 30 or more reviews. The final dataset consisted of information related to 4,490 books, with 610,713 online reviews. Table 5.3a presents the descriptive statistics of our sample.

The numeric ratings for each review were on a 1-star to a 5-star scale where a 1-star corresponded to least satisfied and a 5-star corresponded to most satisfied with

the product. Product sales rank was shown in descending order where a rank of 1 represented the best selling product. Consequently, there was a negative correlation between product sales and sales rank. We used Sales Rank as a proxy for product sales (with the opposite sign).

Table 5.3a Descriptive statistics of sample

Variable	Median	Mean (SD)
<i>ln(Price)</i>	2.41	2.54 (0.58)
<i>ln(SalesRank)</i>	10.11	9.92 (2.00)
<i>AvgRating</i>	4.5	4.18 (0.55)
<i>ln(TotalReviews)</i>	4.01	4.21(0.75)
<i>Helpful votes</i>	2	6.08 (18.75)

We examine the ratings and the writing style of online reviews using the Runs test. Table 5.3b summarizes the items with non-randomness in their reviews. Out of 4,490 books, the sentiment expressed in reviews of 463 books was found to be non-random. The non-randomness could be due to the manipulation of reviews posted by interested parties. It seems that manipulation is less prevalent for the most popular (i.e., sales rank between 1 and 100) and most unpopular books (i.e., sales rank more than 100,000). This indicates that manipulation activity is not affected by the popularity of the book.

Table 5.3b Results of Runs test on randomness of book reviews

	Number of books	Percentage of books with non-randomness in reviews
$1 \leq \text{Sales rank} < 100$	53	9.4%
$101 \leq \text{Sales rank} < 1,000$	292	12.3%
$1,001 \leq \text{Sales rank} < 100,000$	3,076	10.3%
$\text{Sales Rank} > 100,001$	1,069	9.9%
Total	4,490	10.31%

Next, we used a linear regression model to determine if consumers were aware of the manipulations present in the reviews, and if they were able to distinguish between manipulated reviews from non-manipulated reviews. In fact, if consumers were able to differentiate a book review with manipulation from one without manipulation, then with all other information remaining same, a book whose review was being manipulated would either be punished (i.e., resulting in a decrease in sales or an increase in sales rank) or would not be rewarded (i.e., resulting in no change in sales or sales rank). However, if consumers were beguiled by manipulation, then with all the other information remaining same, a book whose review was being manipulated would be rewarded with an increase in sales or a decrease in sales rank.

In the regression model, we examined the impact of manipulation in ratings, sentiments, and readability on the sales rank of the book. Average rating was included as a control variable because previous studies had shown that products with a high average rating enjoyed a high demand. Price and the total number of reviews were included to control for the demand of the book. Given the linear relationship between $\ln(\text{Sales})$ and $\ln(\text{SalesRank})$, we used $\ln(\text{SalesRank})$ as a proxy for sales of books in our log-linear regression models. To control the potential heterogeneity in the existence of manipulation across books with different popularities (as indicated in Table 5.3b), Sales Rank dummies were included in the model as well. Before checking the impact of manipulation on online reviews, we first examined the basic model (Model 1) without the manipulation indices (Equation 5.3.1). Model 2 will be the model that studies the impact of manipulation on sales (as shown in equation 5.3.2.)

$$\begin{aligned} \ln(\text{SalesRank}) = & \gamma_1 \ln(\text{Price}) + \gamma_2 \ln(\text{TotalReviews}) + \gamma_3 (\text{AvgRating}) + \\ & \gamma_4 (\text{sr2_dummy}) + \gamma_5 (\text{sr3_dummy}) + \gamma_6 (\text{sr4_dummy}) + \varepsilon \end{aligned} \quad (5.3.1)$$

$$\begin{aligned} \ln(\text{SalesRank}) = & \beta_1 \ln(\text{Price}) + \beta_2 \ln(\text{TotalReviews}) + \beta_3 (\text{AvgRating}) + \\ & \beta_4 (\text{rating_runs}) + \beta_5 (\text{avg_senti_runs}) + \beta_6 (\text{readability_runs}) + \beta_7 (\text{sr2_dummy}) + \\ & \beta_8 (\text{sr3_dummy}) + \beta_9 (\text{sr4_dummy}) + \varepsilon \end{aligned} \quad (5.3.2)$$

where *Price* denotes the price of each book, *AvgRating* denotes the average consumer rating for each book, *TotalReviews* denotes the total number of reviews for each book, *rating_runs* denotes the runs test result of the rating for each book and is equal to 1 if the test result is non-random, *avg_senti_runs* denotes the runs test result of the average sentiment for each book and is equal to 1 if the test result is non-random, *readability_runs* denotes the runs test result of the readability for each book and is equal to 1 if the test result is non-random, *sr2_dummy* denotes the dummy variable that is equal to 1 for books with sales rank greater than 101 and less than 1000, *sr3_dummy* denotes the dummy variable that is equal to 1 for books with sales rank greater than 1001 and less than 10,000, *sr4_dummy* denotes the dummy variable that is equal to 1 for books with sales rank greater than 10,000. Recall that the product sales rank is shown in a descending order where 1 represented the best selling product. Therefore, the negative correlation between any variable and sales rank indicated that a high value of that variable was associated with higher sales.

Table 5.3c Impact of manipulation on sales

Variable	Model 1 Estimates	Model 2 Estimates
$\ln(\text{Price})$	-0.0254	-0.0254
$\ln(\text{AvgRating})$	-0.1403***	-0.1348***
$\ln(\text{TotalReviews})$	-0.2873***	-0.2905***
<i>Rating_runs</i>		0.0356
<i>avg_senti_runs</i>		-0.2002 ⁺
<i>readability_runs</i>		-0.0439
<i>sr2_dummy</i>	1.2923***	1.2800**
<i>sr3_dummy</i>	4.3210***	4.3057***
<i>sr4_dummy</i>	6.9803***	6.9629***
Intercept	7.0961***	7.1175***
Adjusted R-square	0.6619	0.6619
N	4490	4490

Legend: *** $p < .001$; ** $p < .01$; * $p < .05$; ⁺ $p < .10$

Table 5.3c presents the results obtained using the basic model. We observe that all variables associated with reviews are significantly associated with sales. For example, the coefficient of AvgRating is -0.1403 which indicated that the higher the average rating an item had, the better was its sales (due to the negative correlation between sales rank and sales). Furthermore, the adjusted R-square of the regression model is equal to 0.6619, and it indicates that online reviews can reasonably explain most of the variability in the sales of the books.

Next we studied the impact of review manipulation on sales. The coefficients for *rating_runs*, *avg_senti_runs*, and *readability_runs* captured the incremental impact of manipulation through ratings, sentiments, and readability on sales respectively. We see that the effect of the manipulation of ratings ($\text{para}=0.0356$) and readability ($\text{para}=-0.0439$) on sales rank is not significant. However, on average, the manipulation of sentiments of reviews had a relatively significant impact on sales

rank (para=-0.2001, and p-value<=0.1). This implied that the promotional chat using sentiments in online reviews was effective in generating higher sales for books.

Table 5.3d Impact of manipulation on lagged sales

Variable	Model 1 Estimates	Model 2 Estimates
<i>ln(Price)</i>	-0.02600	-0.0259
<i>ln(AvgRating)</i>	-0.1400***	-0.1325***
<i>ln(TotalReviews)</i>	-0.2860***	-0.2930***
<i>rating_runs</i>		0.01743
<i>avg_senti_runs</i>		-0.2776*
<i>readability_runs</i>		-0.0409
<i>sr2_dummy</i>	1.3588**	1.3433**
<i>sr3_dummy</i>	4.4000***	4.3793***
<i>sr4_dummy</i>	7.03592***	7.0347***
Intercept	7.0123***	7.0567***
Adjusted R-Square	0.6655	0.6658
N	4490	4490

Legend: *** $p < .001$; ** $p < .01$; * $p < .05$; + $p < .10$

Our interpretation for the non-significant results - *rating_runs* and *readability_runs* is that it may be relatively easier for consumers to detect reviews manipulation through ratings or readability, and hence consumers could undo the impact of manipulation of reviews through ratings and readability. The fact that these variables did not generate any significant negative impact on sales might indicate that the consumers were unsure of whether to trust these reviews. As we have explained before, it is indeed hard to differentiate a manipulated review from a review written by zealous customer.

Till now, what we have documented is the correlation between the variables that indicate manipulation of reviews and the sales of books. Next, a time lag is introduced between the dependent variable (measured at time $t+1$) and the variables

representing manipulation (measured at time t) to determine if manipulation at current time influenced the sales of the books in future time. As a result, the baseline model is transformed to Equation 5.3.3:

$$\begin{aligned} \ln(\text{SalesRank})_{t+1} = & \beta_1 \ln(\text{Price})_{t+1} + \beta_2 \ln(\text{TotalReviews})_{t+1} + \beta_3 (\text{AvgRating})_{t+1} + \\ & \beta_4 (\text{rating_runs})_t + \beta_5 (\text{avg_senti_runs})_t + \beta_6 (\text{readability_runs})_t + \beta_7 (\text{sr2_dummy})_{t+1} + \beta_8 (\text{sr3_dummy})_{t+1} + \beta_9 (\text{sr4_dummy})_{t+1} + \varepsilon \end{aligned} \quad (5.3.3)$$

In Equation 5.3.3, we used the notion of relative time rather than absolute time. The assumption is that the manipulation of the early stage reviews would influence the consumers' purchase decisions in the future and result in high sales for the book. For each book, we divided the reviews that it received into two groups, the early stage reviews, and the later stage reviews. The early (later) stage reviews included the first (second) 50% of the reviews that the book received. Then for the early stage reviews of each book (the first 50% of the reviews), we derived the manipulation indices for the variables based on the Runs test. Thereafter, we linked those manipulation indices estimated based on early stage reviews to the sales at the late stage.

In Table 5.3d, the variables *rating_runs*, *avg_senti_runs* and *readability_runs* are based on the early 50% of the reviews for each book. The results shown in Table 5.3d are qualitatively similar to those in Table 5.3c, where the manipulation indices of the variables were built using 100% of the reviews. The effect of manipulation through ratings and readability are still found to be ineffective in the time lagged model. On the other hand, the manipulation using sentiments is found to have a higher and more significant positive impact on sales (para=-0.27764 and p-value<=0.05), which

indicated that vendors are able to influence the final outcomes (in terms of the sales of books) by manipulating the online reviews.

5.4. Implications

Online reviews can be a powerful promotional tool for marketing communication. Marketers and vendors have used this medium because it provides a cheap and impactful channel to reach their customers. In this form of promotional chat or viral marketing, marketers take advantage of networks of influence among customers to inexpensively influence purchase behavior of potential buyers. Reports have shown that promotional chat has infiltrated the online review forums²³. However, it is not clear whether such knowledge sharing sites where customers review products and provide advice to each other are fertile grounds for running promotional campaigns of manipulators. This study examines the extent and the impact of such manipulative actions in the online reviews environment.

In this study, we present a simple but effective way to detect the manipulation of reviews. Our research shows that manipulators use both numeric ratings and textual comment to manipulate online reviews. However, the manipulation of ratings alone is not effective in influencing the sales of books as consumers are able to discover such promotional acts. Instead, manipulation through sentiments is able to significantly influence a consumer's purchase decision. An important benefit of this approach is that one can measure the existence of manipulation in the reviews, and assess the effectiveness of review manipulations in generating sales, without having

²³ <http://www.engadget.com/2009/01/17/belkin-rep-hiring-folks-to-write-fake-reviews-on-amazon/>

access to the backend data about customers' identity that is stored by the host of e-commerce websites.

The use of the Runs test to detect the presence of reviews manipulation through assessment of the randomness of ratings, readability, and sentiments; and then using regression models to assess the effectiveness of online reviews manipulation in generating sales is an important step in discovering the impact of manipulation of reviews. As the method assumes that if the reviews were written by real customers, the writing styles should be random because of diverse background of customers. However, this assumption maybe valid for certain product categories like electronics but not necessarily true for books. Also, it may not be necessarily true for specialty books if customers are of similar background. However, we believe that using the Runs test to detect the manipulated products through assessment of the randomness of ratings, readability, and sentiments, is an important step in discovering the impact of manipulation of reviews. This paper provides a new direction in the detection of online reviews manipulation and its implications.

Chapter 6.

Under-reporting Bias and Online Reviewers' Behavior

User-generated online reviews are a major source of information for movie-goers and can reduce product uncertainty and help consumers infer product quality. Virtually all models for monetizing online UGC, from the well known like Facebook and YouTube to the more obscure like FirstWivesWorld, are based on trust and shared social values (Clemons et al. 2007). The most successful, like the relationship between TripAdvisor and Hotels.com, are based on trust; the greatest failures, like Facebook's Beacon, occur when this trust is violated (Clemons et al. 2007).

Prior research on consumer decision making has established that online reviews are considered more credible and trustworthy by consumers than traditional sources of information (Bickart and Schindler 2001, Li and Bernoff 2008). Despite the subjectivity of online reviews, consumers still pay attention to what has been written

in online reviews to make their purchase decisions (Chatterjee 2001, Chevalier and Mayzlin 2006, Clemons 2008, Clemons and Gao 2008, Dellarocas 2003, Senecal and Nantel 2004). However, to what extent can the evaluations posted by individuals in online networks be considered reliable and representative of the general consensus? This is crucial to understanding the prospects for monetizing customer reviews, and even to their continued relevance in marketing.

Hu et al. (2006) has found evidence that online reviews may not be representative of the general consensus opinions due to under-reporting bias. Under-reporting bias is a form of self-selection bias described in the literature on satisfaction (Anderson 1998). Consumers who are very satisfied or very dissatisfied will be more motivated to voice their opinions through reviews and thus are more likely actually to do so. It has been found that under-reporting bias does exist in certain U.S. online review websites such that the average of reported quality ratings (created by a small population of those sufficiently motivated to post their reviews) do not match the average of perceived quality assessments of the general population. Since consumers are becoming increasingly dependent on online reviews to make purchase decisions, we studied raters'²⁴ behaviors to reveal whether under-reporting bias exists across cultures, and whether online consumer rating behavior will yield biased or unbiased estimators of a product's quality in various markets.

Since each posted online review is an assessment of a single individual's perceived quality of a product, this study first explores how such reported quality could be influenced by cultural factors. Siau et al (2010) finds that that national

²⁴ In this study, we will be using the terms 'rater', 'reviewer', 'poster' and 'consumer' interchangeably.

culture has an impact on knowledge sharing in virtual communities. Thus, it is anticipated that the behavior of individuals in online networks may also across cultures, and may differ from offline behavior as well.

Behavioral theory in social psychology asserts that specific salient beliefs influence behavioral intentions and subsequent behavior (Ajzen 1985, 1988, 1991). Employing constructs from behavioral theory – attitude, social norms and motivation — we seek to understand the following important questions, which to the best of our knowledge have not been answered in previous online review literatures:

- What factors motivate consumers to write online reviews?
- How does culture influence raters' behavior when writing reviews and how do cultural differences manifest in differences among ratings?

To identify the potential under-reporting bias that might render the mean of online movie reviews a biased estimator of movie quality; we compare the distribution of voluntarily posted online movie reviews to those reviews which we believe are closer to the distributions of true perceived quality. In this regard, we conduct a survey in which respondents were asked to report their ratings for a number of movies they have viewed and under what circumstances would they be more likely to post online reviews. Comparing survey results to posted online reviews in each cultural environment, the results show that under-reporting bias varies across different cultures - online reviews reflect a movie's perceived quality in Chinese online networks more accurately than in the U.S.

To understand the behavior of movie raters from different cultures, we draw upon some of the behavioral theory on attitude, social norms, and under-reporting bias and examine how their behaviors are influenced by cultural differences.

Attitude and Social Norms

Taylor and Todd (1995) describe a construct that argues that behavioral beliefs influence attitudes, which in turn determine intentions and actual behavior. Behavioral beliefs arising from social pressure are termed normative beliefs (Ajzen 1991), also termed *social norms*, which is the influence created by a person's normative beliefs that others approve or disapprove a particular behavior. People's intentions to perform a particular action are influenced by social norms, or by their perception that important others think they ought to perform those actions. In our context, social norms refer to the influence from consumers' normative belief that the behavior is accepted, encouraged, and promoted by their social circle. Consumers may believe that their family, friends, and even online peers would favor certain online opinions, and this belief tends to influence their intentions and opinions. We examine how offline interactions and social norms influence online social network behavior.

Cultural Differences

Hofstede's (1980) cultural dimensions serve as the most influential theory of culture and cultural differences in research in the social sciences (Nakata and Sivakumar 2001); his categorization of national societies is also widely used as the basis of applied research in the study of marketing differences across cultures and in e-commerce studies (Pavlou and Lin 2002). His cultural framework has also received

strong empirical support (Sondergaard 1994). The framework was generated through the most extensive examination of cross-national values ever undertaken, involving 116,000 respondents from 40 countries (Pavlou and Lin 2002). The results were consistent with the findings in 38 other studies (Nakata and Sivakumar 2001). Hofstede separated cultures on the basis of (a) masculinity-femininity, (b) individualism-collectivism, (c) power distance, (d) uncertainty avoidance, and the recent addition of the Confucian dimension of (e) long-term orientation (Hofstede 2001). Our work starts by accepting Hofstede's framework; we are not testing it for validity, but attempting to demonstrate whether the behavior of online raters is consistent with this theory. We focus on the implications of dimensions (b) individualism-collectivism and (e) long-term orientation.

Individualism-collectivism refers to the basic level of behavior regulation of either individuals or groups. Individualists view self and immediate family as relatively more important than the collective. *Long-term orientation* as described by Hofstede suggests following tradition, perseverance and the practice of benevolence; *short-term orientation* on the other hand, is the tendency towards consumption and materialism. As these are long-established and influential theories of culture and cultural differences, we will be using these cultural constructs in our conceptual development to better understand how different forms of national "culture" manifest themselves in online interactions (Ess and Sudweeks 2005). We caution the reader not to view these dimensions as merely cultural stereotypes. Hofstede is not suggesting that all Chinese are benevolent towards all other humans, or even towards all other Chinese, in their online behavior, nor is he suggesting that Western culture is

without benevolence and the Golden Rule of “Do unto others” Hofstede is suggesting that with a large enough sample, differences in cultural norms are readily observable.

Under-Reporting Biases

Hu et al. (2006) found evidence of two self-selection biases, acquisition bias and under-reporting bias, in the reporting of online consumer reviews, both of which render mean ratings a biased estimator of product quality. Acquisition bias refers to the situation that only consumers with a favorable disposition towards a product will acquire the product. Since only consumers with a pre-acquisition utility perception higher than the product’s posted price are willing to pay the price to acquire, and thus have the chance to review the product, this creates a bias towards a greater number of positive product reviews. Secondly, consumers who are greatly satisfied or greatly dissatisfied are more likely to report their review; correspondingly, those consumers with more moderate sentiments are less likely to post a review. This is termed *under-reporting bias*.

Based on the data collected from Amazon.com and an offline survey conducted on U.S. customers, Hu et al. (2006) documented that while online consumer reviews have a J-shaped distribution, actual consumer assessments for the same set of products are normally distributed. They concluded that online reviews do not reflect a product’s perceived quality across the population of all users, which they term its *true perceived quality*. Rather, online reviews quite naturally reflect the views of those who post them, which differ from true perceived quality because of under-reporting by those customers with moderate views.

However, their study did not examine the raters' attitude and social norms across different cultures. Since the degree of under-reporting bias might vary across cultures, we set out to understand raters' behavior across different cultures — American and Chinese — and to identify under what circumstances online reviews might or might not reflect a product's true perceived quality in different settings. Since a large proportion of the movies watched by Chinese consumers were downloaded at no cost, acquisition bias is not very significant in our context. Hence, we focused on under-reporting bias for this study.

The United States and China were chosen for this study because they represent almost reverse positions on several important cultural dimensions (Hofstede 1980). In addition, we chose to collect data on Singapore because of its mixture of Western and Eastern culture, which allows us to see if and how culture mediates the attitude and behavior in such a hybrid culture.

The next section describes the theoretical framework and our research hypotheses. The research setting and methodology are presented in Section 6.2. Section 6.3 and 6.4 provides the analysis of our empirical findings. Section 6.5 presents the survey results and examines under-reporting bias. Section 6.6 provides a robustness check on our findings. Section 6.7 discusses the limitations and concludes with suggestions for future research.

6.1. Conceptual Development

The proposed research model of online movie reviewers' behavior is adapted from Pavlou and Lin (2002). The dependent variable – online review behavior, as

measured by the rating assessment each reviewer gives to each movie – captures consumers’ reviewing behaviors. Drawing from behavioral theory in social psychology, two factors that directly influence reviewers’ intentions towards the reviewing process included in this study are attitudes towards the movie and social norms regarding what is customary in reviews (Ajzen 1991, Pavlou 2002). We investigate the relationships among these in terms of cultural differences, using the dimensions of individualism / collectivism and long-term / short-term orientation. In addition, we gathered survey data that enabled us to look at the motivation for consumers to write online reviews.

6.1.1. Attitude

Attitude has been used as a predictive factor that influences behavioral intention in multiple theories, such as the *theory of planned behavior* (TPB) (Ajzen 1988) and the *theory of reasoned action* (TRA) (Fishbein and Ajzen 1975). These theories have gained substantial empirical support (Madden et al. 1992, Pavlou and Lin 2002). Attitude here refers to an overall evaluation of the movie that an individual has viewed. A favourable attitude towards a movie will positively influence the rating of online movie reviews. Of course, it is not surprising that a reviewer’s attitude towards a movie, which is essentially his or her assessment of that movie, will be a prime determinant of the content of the review posted. But it is not the only determinant.

6.1.2. Social Norms

Social influence is related to Hofstede’s dimension of individualism / collectivism, and is a second factor that directly influences online reviews.

Collectivism refers to the extent to which individuals feel themselves to be integrated into groups and the extent to which opinions are informed by group norms and expectations or even formed based on these norms and expectations (Hofstede and Bond 1988). Members of individualistic societies are more likely to value freedom of expression, while those of collectivistic culture are more likely to seek to group consensus. China has for centuries been highly collectivist; in particular, Chinese attention to group norms predates collectivism in the sense imposed by Communism and indeed goes back to China's Confucian heritage. Conversely, the United States is among the most highly individualistic societies. Consequently, we expect there to be differences in the effect of societal influence on individual behavior, and specifically for this study we expect to be able to observe these differences by comparing online movie reviews contributed by members of the two cultures.

Collectivist societies have strong relations within the extended family and among friends and acquaintances (Hofstede and Bond 1988). Their group relations seek to maintain harmony by going along with the group's wishes and by promoting and maintaining long-term relationships (Bond and Smith 1996). We anticipate that members of a collectivist culture, such as China, would want to maintain harmonious relationships among participants, both as readers and as writers, in the online movie review website. On the other hand, we expect that U.S. movie reviewers value freedom of expression more strongly and hence feel themselves to be freer to openly express their appreciation or great dissatisfaction of the movies they have viewed. Indeed, as noted by Jaron Lanier (Lanier, 2010) in his recent book, the anonymity

made possible by websites seems to encourage the emergence of Internet trolls in the west and a practice he calls drive-by anonymous insults.

Due to their recent colonial history, Singaporeans have been influenced by Western culture, but because of their earlier history, Eastern culture and values are also strong. Values such as obedience and harmony are important, and they value intense friendships and trust within the family. Therefore, we anticipate that Singapore is more of a collectivist society than the United States. Since China is highly collectivistic and the U.S. is highly individualistic, we expect that the examination of review patterns from China, Singapore, and the United States will reveal significant differences in reviewer behavior, consistent with Hofstede's cultural classification. We anticipate that the ratings posted by American reviewers will more clearly express their likes and dislikes for movies. On the other hand, the ratings given by Chinese reviewers will be more constrained and more narrowly confined within a tight range centered on the average of the ratings given by previous customers. Thus, we expect that an attitudinal difference in the reviewing intentions among the three countries, again consistent with Hofstede, as the following hypotheses propose.

***Hypothesis 1:** Collectivist societies tend to place greater focus on harmony and thus tend to write fewer extremely negative reviews than individualist societies.*

***Hypothesis 2:** The value placed upon freedom of expression is reflected more in the ratings of movies from individualist societies than those from collectivist societies.*

***Hypothesis 3:** Societal norms have greater effect on the ratings of movies in collectivist societies than in individualist societies.*

According to Hofstede (2001), China is ranked extremely high on the dimension of long-term orientation, which reflects the impact of the teachings of Confucius on Chinese culture and society. One of the key principles of Confucian teaching is the basic human benevolence toward others and this consists of treating others as one would like to be treated²⁵. We therefore expect Chinese to be less willing to fully express their dislike in their ratings for bottom-ranked movies than American reviewers would be. (Notice, we cannot distinguish whether Chinese have more generous views of the movies, or merely restrict themselves to more generous public statements and posted reviews, solely on the basis of the posted reviews. Additional hypotheses address these differences, and our methods for studying these differences are described in the section on research methodology. But, regardless of the motivation, just as we would expect Chinese reviewers to be more generous *in general* (as expressed in hypothesis 1), we would expect them to be more generous *even in the case of the worst and most disappointing experiences* (as expressed in hypothesis 4 below).

Hypothesis 4: The ratings for Bottom-ranked movies given by collectivist societies will be less extreme than in individualist societies.

We have applied Hofstede's (2001) cultural classifications, allowing us to predict certain differences in behaviors across cultures, and our hypothesis allow us to

²⁵ While there is no immediately obvious connection to a Western observer between benevolence and a long-term orientation, historically Confucian teachings have stressed long-term orientation, collectivist ties to family and society, and a higher degree of benevolence within groups. Likewise, Hofstede describes group averages; at no point does he suggest that absence of benevolence in individualistic societies, or the absence of altruistic behavior in the West. Explicitly, the Judeo-Christian traditions of the West do acknowledge the importance of treating others as you would wish to be treated; still, Hofstede expects to see a greater degree of benevolence in the West, and, assuming this is true, we would expect to see differences between Chinese and American online movie reviewing.

analyze the extent to which the predicted differences do or do not appear as in the specific context of online behavior in the specific domain of online movie reviewing. We are not attempting to test Hofstede's theory, and we do not argue that all Chinese are collectivist, or that all Westerners are extreme individualists. We simply analyze millions of movie reviews posted on websites that cater primarily to Chinese or to American reviewers, and look for the predicted differences in rating behavior. The differences are indeed consistent with the predictions based on Hofstede's theory. We understand that not all movie-goers rate the movies they have seen; significantly, the absence of ratings follows a pattern, with more extreme under-reporting bias in the United States than in China, which is consistent with our hypotheses and with Hofstede's cultural classification upon which they are based.

6.1.3. Motivation

Writing reviews seems to address basic human needs for belonging to and gaining acceptance from groups in which they participate; and for achieving status and recognition (Maslow 1943). We hypothesize that consumers who participate in the writing of online reviews are motivated to meet these needs. If this is true, then writing reviews would be based both on individual motivations and on the interaction of these motivations with social norms. An individual from a highly collectivist society would most definitely not achieve his desire to feel as if he were part of a group if his reviews violated the norms of the group, and would not receive self-esteem and recognition if his reviews were rejected because they violated the norms of the group. Thus, we assert the following two hypotheses:

Hypothesis 5: *The motivation to write movie review is affected by people's social needs to feel a sense of belonging and sharing, which may require that reviews adhere to social norms.*

Hypotheses 6: *The motivation to write movie review is affected by people's needs for esteem and recognition, which may require that reviews adhere to social norms.*

6.2. Research Methodology

Our research methodology involves three specific sets of analyses. First, we compare the rating behavior of Chinese and American reviewer using data collected from Douban.com and IMDB.com. Second, we perform attitudinal studies to determine, to what extent do online reviews reflect a product's true perceived quality. In addition, we study how likely a U.S. or Chinese reviewer will be affected by the reviews that were posted previously. Finally, we investigate to what extent Singapore movie raters resemble those of China and to what extent they resemble those of the United States.

6.2.1. Cross-Cultural Data

To study the cultural differences in online movie review behavior, we gathered the reviews from two online movie review websites, IMDB.com and Douban.com. IMDB.com was chosen because it is the largest online movie review website with over 57 million visitors each month. For the Chinese website, we chose Douban.com because it is a cloned version of IMDB in China, and is consistently ranked as one of the most popular online review website in China by Alexa Internet (2008). To ensure

we are comparing members of the U.S. and Chinese cultural communities, we only crawled the ratings from IMDB that were posted by U.S. reviewers.²⁶

Our data sample contains two datasets. For the first dataset, reviews were collected on 1,000 movies randomly selected from IMDB and 1,000 movies randomly selected from Douban, using a random counter on the movie identification number. While we expect that 1,000 movies is a well representation of the movies across the two websites, it is possible that our results might be influenced by the movies selected from these websites. To control this, we conducted experiments with a second dataset. For this dataset, we first chose the Top 100 and Bottom 100 ranked movies in IMDB. Then based on these movie titles, we collected the same movies titles in Douban.com and their corresponding movie reviews from the two sites. By using the same movies to compare the rating pattern, we attempt to ensure that the observed differences in reviews are due to inherent differences in reviewer behavior rather than differences in the movies selected for comparison. The reason for focusing on those top and bottom-ranked movies is that if indeed under-reporting bias were present, we believed that it would be more likely to be observed for movies within such categories.

Each dataset has its own advantages and disadvantages. Using 1,000 movies selected separately and at random from these two movies reviewing sites reveal in

²⁶ Each movie has a webpage that shows the ratings given by U.S. raters. For instance the ratings given by U.S. users for the movie “The Godfather” can be crawled from <http://www.imdb.com/title/tt0068646/ratings-usa>. Nevertheless, the fact that a reviewer resides in the U.S. does not necessarily mean that the reviewer is an American, or that the reviewer has adopted the behaviors that Hofstede typically associates with an individualist culture. However, to the extent that our data set might include Chinese residing in the United States, this inclusion actually should lessen the strength of the effects we were measuring. Thus, our results are a conservative test and may actually understate the effect that we have claimed.

general how movie raters' behavior is different across these two websites. However, this introduces the possibility that the observed difference is driven by the different movie titles selected from the two sites. Using the top and bottom-ranked movies and then comparing reviews from the two sites eliminates the above possibility, but introduces the possibility that audience response to different movies might be different between the two cultures. For robustness check, we also collected data from the Top 100 and Bottom 100 ranked movies in Douban, and the corresponding movie titles and their respective movie reviews in IMDB. This will be presented in the robustness check Section 6.4.

Table 6.2.1a: Statistics of Dataset from IMDB.com

Category	Number of Movies	Number of Ratings
Top 250	(top) 100	2,944,037
Bottom 100	100	191,411
Entire Collection	(random) 1000	691,739

Table 6.2.1b: Statistics of Dataset from Douban.com

Category	Number of Movies	Number of Ratings
Top 250	(top) 100	483,680
Bottom 100	100	4,151
Entire Collection	(random) 1,000	14,645,654

The summary statistics of the dataset of the second batch from both websites are shown in Table 6.2.1a and 6.2.1b. Data collection for both batches started on 20th December 2008 and ended on 15th January 2009.²⁷ For each item, we collected the movie title, movie ID, and review information. Specifically, for each movie review, we gathered the numeric rating, review date and the original text of the review. On

²⁷ The data analysis is based on data collected prior to 15 January 2009. The collection of dataset by the random process was slow because we were constantly blocked by IMDB and Douban.

Douban, consumers can report an integer movie review on a 5-point Likert-type scale, anchored at 1-star = least satisfied and 5-star = most satisfied. On IMDB, consumers can report an integer movie review on a 10-point Likert-type scale, anchored at 1-star = least satisfied and 10-star = most satisfied.

6.2.2. Experimental Calibration with Survey Data

After comparing the rating distribution of Chinese consumers to that of U.S. consumers, we conducted a survey in which respondents were asked to review the movies they have viewed. Then we compared the survey results with those observed on IMDB and Douban. We expected that this survey mechanism would result in more balanced reviews with less under-reporting bias, and thus in reviews ratings that more accurately reflect the community's perceived average quality for movies in our sample. This is essential to explore our hypotheses about cultural differences in reviewing behavior.

To gather the survey from Chinese and Singaporean, our respondents were university students attending business, information systems and economic courses. Each student was asked to review 16 movies that vary in terms of category and genre. For each movie, the subjects were asked to rate the movies they have viewed and report their intention and motivation to write online movie reviews. The online survey instrument was emailed to 1,500 students composed of native Chinese and Singaporeans who spoke Chinese fluently. Invitation emails explained the purpose of the study and requested participation. To gather the survey response of U.S. participants, an invitation request was posted on Facebook. For both email and Facebook requests, respondents who clicked on the URL link provided in the email

message (or on Facebook) were then directed to a website to take the online survey, of which 87 Chinese students, 212 Singaporeans and 247 Americans responded. Participation was voluntary and the response rate for the email invitation was approximately 20%. The movie rating scales for the survey were based on those of Douban.com. A preliminary version of the survey was generated and reviewed by doctoral students for clarity. Finally, to verify the appropriateness of the survey, it was pre-tested with multiple research students who varied with age, gender and education. Since this attitudinal survey involved U.S., Chinese and Singaporean nationals, we are able to assess the extent to which Singaporeans do or do not differ from American / Chinese in terms of their rating behaviors.

6.2.3. Graphical Data Analysis

We retrieved movie reviews of 1,000 randomly selected movie titles from both Douban and IMDB, and we focused on movies with an average review of 3-stars in Douban and those with 5-stars in IMDB, the median ratings for the two sites respectively. After that, we plotted the distribution of the ratings from Douban and from IMDB (as shown in Figure 6.1 and Figure 6.2). Theoretically movies with these median average ratings are more likely to be normally distributed, therefore, we expect to observe a normal distribution for both but as is evident from even a quick visual inspection of Figure 6.1 and 6.2 we did not. The behaviors of the two populations indeed do differ.

Figures 6.1 and 6.2 show that for all movies with average review score equal to the median (92 movies on IMDB and 151 on Douban), the rating histogram for IMDB is W-shaped, whereas for Douban, the histogram is indeed normal and bell-

shaped. Thus there are differences in the rating pattern even for average-ranked movies. To ensure that our results are not driven by the efforts of reviewers, we checked if reviews with only ratings differ from reviews with both ratings and text comments. Since preparing a review is more time-consuming than merely providing a numeric evaluation and might suggest a more serious effort to assess accurately, we have compared the numeric ratings of reviews with and without textual reviews across IMDB and Douban and found them to be similar (Wu and Huberman 2007). Our results show that the rating histogram for Douban is bell-shaped, while that of IMDB is W-shaped.

In the work that follows, we compare individual rating behavior using the second dataset, in which we ensure the ratings are for the same set of movies. Since the movies being compared are the same, any observed differences in the rating patterns are driven by the cultural differences between the rating populations, and not by the selection of movies included in each sample. We focused on ratings of the most extreme movies, the top-ranked and bottom-ranked movies, in order to examine the most extreme rating behavior. Unless we state otherwise, all subsequent analysis was done using the second dataset.

Figures 6.3 and 6.4 show the rating distributions for the top-ranked movies in IMDB and the corresponding rating distribution for the same set of movies in Douban. Figure 6.3 once again displayed the characteristic W-shaped distribution that we previously observed among U.S. raters, whereas Figure 6.4 once again has a unimodal distribution. It seems that on average, Chinese reviewers are more reserved in giving the highest ratings than are U.S. reviewers; they are also more reluctant to

assign the most negative reviews. From both figures, we see that on the whole, the rating behavior of reviewers of top-ranked movies were not very different between both cultures, except for some modest limitation of top and bottom reviews among Chinese reviewers.

Figures 6.5 and 6.6 show the distribution of the ratings for the Bottom 100 movies in IMDB and the corresponding rating distribution for the same set of movies in Douban respectively. Unlike the results for the top-ranked movies, there is a great difference between Chinese and American reviews for IMDB bottom-ranked movies. Figure 6.5 shows that for the U.S. reviewers, the largest number of reviewers gave a very low rating of 1-star, resulting in a U-shaped distribution. For the Chinese reviewers, even when a movie is bad, the online reviews still demonstrate a bell-shaped distribution, as shown in Figure 6.6.²⁸ This supports Hypothesis 4, which argues that the ratings for bad movies given in collectivist societies will be less extreme than individualist societies. U.S. reviewers on the other hand are more inclined to express their dissatisfaction vigorously and openly, and often in the most extreme terms, which is consistent with Hypothesis 2. This probably partly explains the difference in the number of reviews for the Bottom 100 movie category in IMDB (191,411 total ratings) and Douban (4,151 total ratings), providing support for Hypothesis 1. (A competing explanation is that the very worst movies screened in America are unlikely to be watched by Chinese audiences, which may also explain some of the observed difference in the counts of reviews for the worst ranked movies.)

²⁸ To check the consistency of this result, we conducted the same analysis based on independent datasets collected prior to 15 September and prior to 15 December 2008. We still find the W-shaped distribution for the U.S. and a bell-shaped distribution for the Chinese.

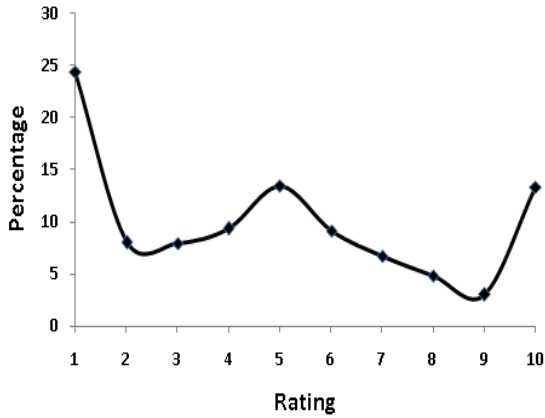


Figure 6.1: IMDB Movies with Avg Rating= 5 (92 out of 1000 movie items)

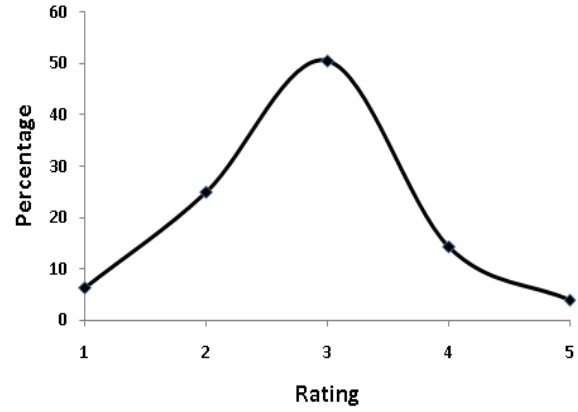


Figure 6.2: Douban Movies with Avg Rating = 3 (151 out of 1000 movie items)

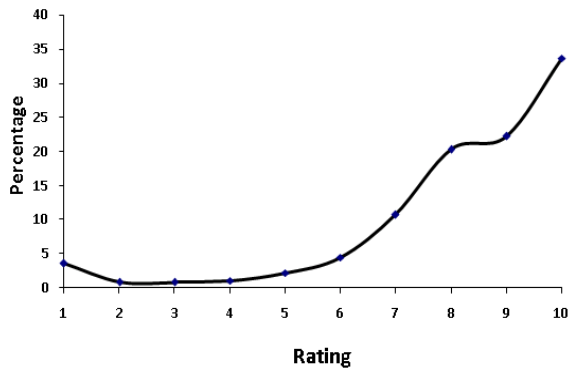


Figure 6.3: IMDB Top-Ranked Movie Ratings

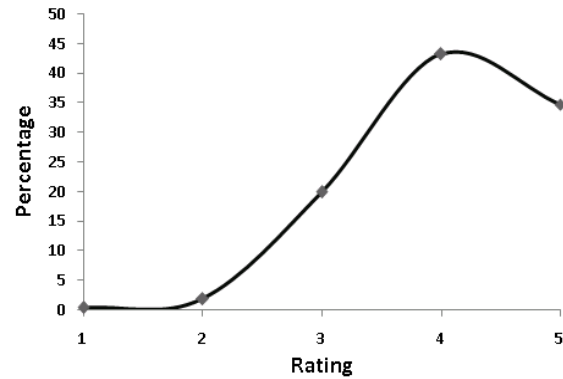


Figure 6.4: Corresponding (Top-Ranked) Movie Ratings in Douban

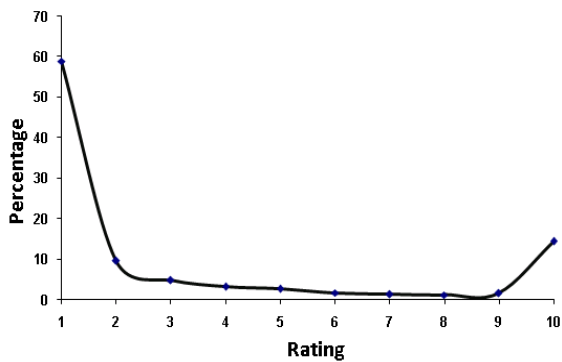


Figure 6.5: IMDB Bottom-Ranked Movie Ratings

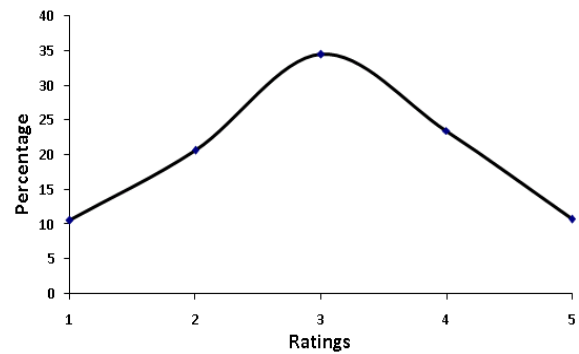


Figure 6.6: Corresponding (Bottom-Ranked) Movie Ratings in Douban

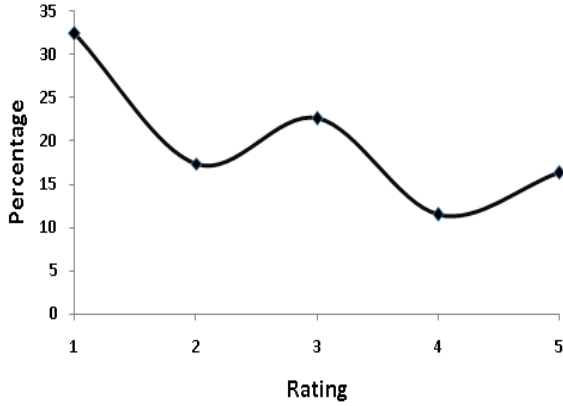


Figure 6.7: IMDB Movies with Average Rating (92 out of 1000 movie items) on same scale

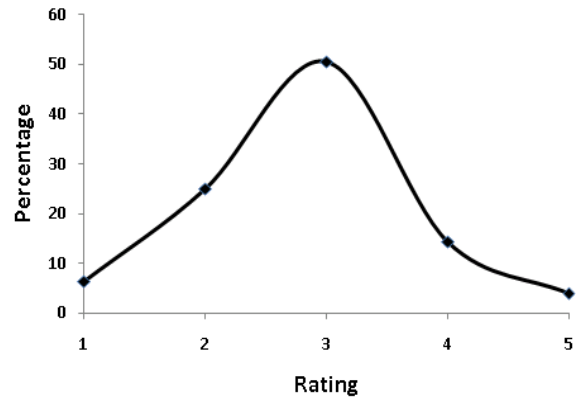


Figure 6.8: Douban Movies with Average Rating (151 out of 1000 movie items) on same scale

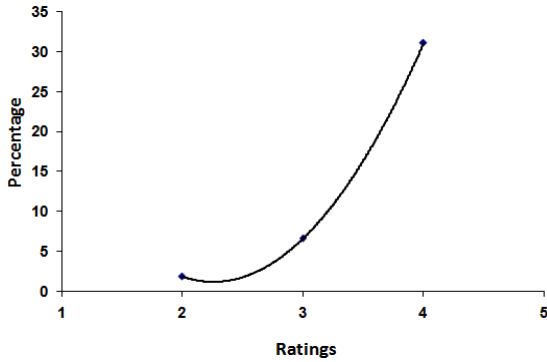


Figure 6.9: IMDB Top-Ranked without Extremes

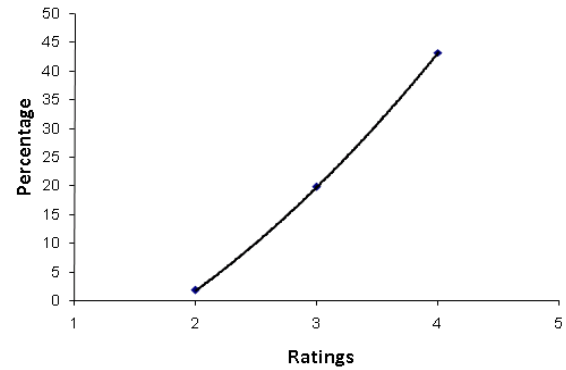


Figure 6.10: Corresponding Top-Ranked without Extremes in Douban

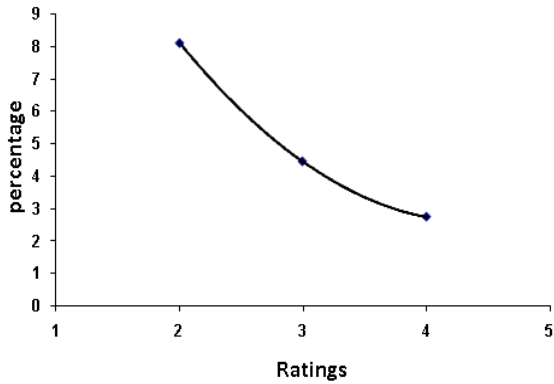


Figure 6.11: IMDB Bottom-Ranked without Extremes

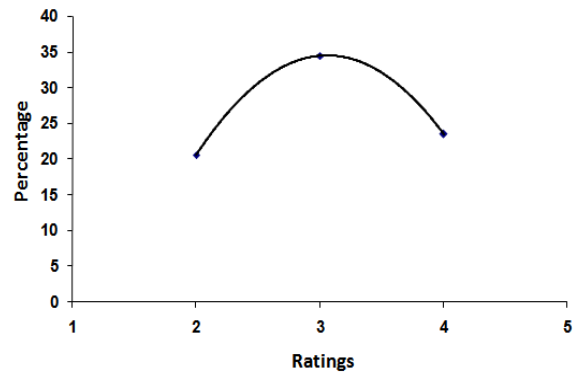


Figure 6.12: Corresponding Bottom-Ranked without Extremes in Douban

Prior studies (e.g., Tourangeau et al. 2000, Poulton 1989), have reported that respondents tend to avoid the extremes in surveys. In psychophysics, this trend is

called response contraction bias (Poulton, 1989). However, this is not what we have observed in IMDB.com; there are in fact more responses at the extreme endpoints of 1s and 10s, this is the case even when we convert the results to the same scale as shown in Figures 6.7 and 6.8. For Figures 6.3 to 6.6, if we first convert the 10-point IMDB scale to a 5-point scale, consistent with Douban, and then remove the extreme ratings — the ratings of 1 and 5 — this will result in Figures 6.9 to 6.12 respectively. The results suggest that the extremes are over-represented in the American rater population relative to the Chinese rater population, and perhaps the extremes in the American rater population are even over-represented relative to the American population of movie-goers more generally. There are several explanations. Perhaps Americans are more honest and willing to post extreme views because they are less influenced by the mean. Another explanation is that Americans might try to be different by giving extreme ratings, since simply giving an average rating does not show that they are individuals. Alternatively, Americans may be less willing even to rate unless extremely motivated by very strong attitudes, positive or negative, towards the film. The Chinese, on the other hand are demonstrably less likely to give extreme ratings, perhaps because they are more influenced by the consensus and the average sentiment of the reviews already posted.

Overall there is a huge difference between consumer reviews of all the movies in IMDB and Douban. This difference is visible when comparing reviews of all movies, that is, when comparing Figures 6.7 and 6.8. But it is most visible and most pronounced when comparing review of the bottom-ranked movies, that is when comparing Figures 6.11 and 6.12. It seems that American reviewers with the most

extreme opinions are greatly over-represented relative to the population at large, which distorts American movie ratings relative to the true perceived quality. These results lead us to believe that there is less under-reporting bias among Chinese reviewers, as shown by the bell-shaped curve in Figure 6.8. We were concerned that this effect could also be explained not by under-reporting bias, but by profoundly different responses to the movies when seen by Chinese and American audiences, in which case Chinese viewers might have far more average assessments than American audiences, rather than different propensity to review based on the strength of their assessments. To address this concern, we conducted a survey to compare the distribution of assessments of the two populations, to verify whether the difference in reviewing behavior was due to different levels of under-reporting bias or to some other, perhaps as yet unreported, behavioral difference.

6.3. Impact of Attitude and Social Norms on Rating Behavior

To examine the proposed hypotheses, the first experiment was designed to study how ratings are influenced by general consensus:

Rating = f (*Average rating*, *Average rating in IMDB Bottom movies*, *Average rating in IMDB Top movies*, *Average rating in Douban Top movies*, t)

This is translated into the following empirical model:

$$\begin{aligned}
 Rating_{jt} = & \alpha_1 * AvgRating_{j,t-1} + \alpha_2 * AvgRating_{j,t-1} * IMDB_Bottom + \alpha_3 * AvgRating_{j,t-1} * IMDB_Top \\
 & + \alpha_4 * AvgRating_{j,t-1} * Douban_Top + \alpha_5 * t + \alpha_6 * Bottom_Dummy \\
 & + \alpha_7 * Imdb_Dummy + \varepsilon_{jt}
 \end{aligned}
 \tag{6.2.4}$$

where:

t	the sequence order of each review to control the temporal effect. The first review posted for a movie will have $t = 1$.
$AvgRating_{j,t-1}$	denotes the average consumer rating at the time when the $(t-1)^{th}$ review was written for movie item j .
$IMDB_Bottom$	IMDB_Bottom is a dummy variable and equals 1 for the ratings of Bottom movies in IMDB.
$IMDB_Top$	IMDB_Top is a dummy variable and equals 1 for the ratings of Top movies in IMDB.
$Douban_Top$	Douban_Top is a dummy variable and equals 1 for the ratings of Top movies in Douban.
$Bottom_Dummy$	Bottom_Dummy is a dummy variable and equals 1 for movies in the Bottom category.
$IMDB_Dummy$	IMDB_Dummy is a dummy variable and equals 1 for movies in the IMDB category.

We estimated Model 1 using robust regression procedure²⁹ (Yaffee 2002, Chen 2002) to study how ratings are influenced by online review environment variables and presented the results in Table 6.2.4. None of the between variable correlations is larger than 80% and the condition index is about 35. A further look at this condition index shows that it is related to the correlation between the main effect and the interaction terms. Hence multicollinearity is not much of concern in our analyses. If we consider a boundary of -1 and 1, our results show that for both cultures, raters have different rating tendency for the top and bottom-ranked movies. Particularly, the parameter estimate for average rating is significantly positive ($AvgRating = 0.9279$ and $p\text{-value} \leq 0.001$). This means that when Chinese consumers rate a bad product (in this case a movie), if the consensus rating is -1, new consumers will give a rating of -0.9279, a slightly less negative score that is within the boundary of -1 and 1. The

²⁹ We used the PROC ROBUSTREG procedure in SAS, which attempts to down-weight outlying observations and calculate stable and resistant estimators using robust regression techniques. This solution addresses the non-normality and heteroskedasticity issues created by outliers (Yaffee 2002, Chen 2002).

same principle applies when Chinese reviewers rate a good movie; if the consensus rating is 1, new reviewers will give a rating of only about 0.24, which is the sum of 0.9279 and -0.6917. This, again, is a less positive score with tendency to rate towards the center. Overall, the results show that Chinese reviewers tend to leave ratings within the boundary of the general consensus and are not likely to post a rating that is more extreme than the average of what the community has already given. This provides support for Hypothesis 3.

Table 6.2.4: Average ratings in IMDB and Douban

Variable	Model
<i>AvgRating</i>	0.9279***
<i>Avg_Rating * IMDB_Bottom</i>	0.2631*
<i>AvgRating * IMDB_Top</i>	0.1738
<i>AvgRating * Douban_Top</i>	-0.6917***
<i>Intercept</i>	3.4503***
<i>Sequence</i>	-0.0033***
<i>IMDB dummy</i>	1.6835***
<i>Bottom dummy</i>	-3.3391***

Legend: *** $p < .01$; ** $p < .05$; * $p < .10$

However, for the U.S. reviewers, the story is different, especially when they are facing a movie that they perceive to be of low quality. The interaction between *IMDB_Bottom* and average rating is positive ($AvgRating * IMDB_Bottom = 0.2631$ and $p\text{-value} \leq 0.001$). This indicates that when American consumers rate a bad movie, if the consensus rating is -1, new consumers will give a rating of -1.19 (which is the sum of -0.9279 and -0.2631) a more negative score with tendency to move out of the boundary of -1 and 1. Overall, it seems that U.S. reviewers are not confined to the rating boundary of the community, and they are more willing to post extreme

reviews, especially when they are dissatisfied. Perhaps Americans value the need to clearly express their likes and dislikes for movies are less influenced by the mean, which again gives support for Hypothesis 2.

6.4. Motivation for Writing Online Reviews

For this section, we examine what motivates consumers to post online movie reviews. In our survey, we asked the respondents:

- (1) If they have been to movie review websites?
- (2) If they have ever rated a movie in movie review websites?
- (3) Under what circumstances would they rate a movie online —
 - (a) when they like the movie?
 - (b) when they dislike the movie or
 - (c) when they want to share their opinions with others?
- (4) How much influence do online movie reviews have on them?
- (5) How often do they rate movies?

From the survey results collected from our Chinese and American respondents, we perform a logistic regression³⁰ to study the motivation for consumers to write online reviews. Our dependent variable is the frequency of respondents rating movies

³⁰ We also conduct our analysis using Ordinary Least Square and robust regression procedure. The results are qualitatively similar.

(Question 5), where 1 means ‘never rated’ and 5 represents ‘often rated’.³¹ Our dependent variables capture the answers for question 1 to 4³²:

$$\text{Frequency} = f(\text{Been_to}, \text{Rated}, \text{Like}, \text{Dislike}, \text{Share}, \text{Influence})$$

Table 5.2.2a: Motivation to post ratings for Chinese

Variable	Model
<i>Been_to</i>	1.5663**
<i>Rated</i>	1.1396***
<i>Like</i>	0.6130**
<i>Dislike</i>	0.0604
<i>Share</i>	0.6647**
<i>Influence</i>	0.0331
Intercept	-3.6553***

Legend: *** $p < .01$; ** $p < .05$; * $p < .10$

Table 5.2.2b: Motivation to post ratings for American

Variable	Model
<i>Been_to</i>	1.5659**
<i>Rated</i>	1.2525***
<i>Like</i>	-0.5994
<i>Dislike</i>	-0.6002
<i>Share</i>	-0.3486
<i>Influence</i>	0.1357
Intercept	-2.2938**

Legend: *** $p < .01$; ** $p < .05$; * $p < .10$

In Table 5.2.2a shows the results on the motivation for Chinese to post ratings, the variables of “*Been_to*” and “*Rated*” are control variables, while the remaining are

³¹ Our dependent variable is on the Likert scale of 1 to 5 where 1 means ‘never rated’, 2 means ‘rated once’, 3 means ‘seldom’, 4 means ‘sometimes’ and 5 means ‘often rated’.

³² Question 1 is for the first dependent variable “*Been_to*”; Question 2 for the second dependent variable “*Rated*”; Question 3a for the third dependent variable “*Like*”; Question 3b for the fourth dependent variable “*Dislike*”; Question 3c for the fifth dependent variable “*Share*” and Question 4 for the sixth dependent variable “*Influence*”.

our variables of interest. We see that the desire to share opinion with others (*Share* = 0.6647 and $p\text{-value} \leq 0.05$) dominates all other expressed reasons for posting reviews in motivating a consumer to write a review, giving strong support for Hypothesis 5, at least among reviewers from collectivist culture. The next factor that motivates consumers to write online movie reviews is the desire to express their liking for the movie (*Like* = 0.6130 and $p\text{-value} \leq 0.05$). Consistent with Hofstede's cultural classifications and with our hypotheses, when Chinese raters like a movie, they are more inclined to write online reviews; hence the much smaller numbers of truly negative online reviews posted on Douban. Finally, individuals from collectivist culture do not appear to write online reviews because of the need for esteem or to influence the views of others, rejecting Hypothesis 6. To summarize, consumers from collectivist culture are more likely to speak out when they like a movie instead of when they dislike a movie.

On the other hand, the motivation for American to post ratings is rather bleak. Table 5.2.2b shows insignificant results for all variables of interest. The posters of reviews in the U.S. culture do not appear to write reviews because they like or dislike the movie, nor that they want to share their opinions or influence others. Americans who are motivated to post reviews seems to be driven by other unknown reasons. Perhaps an interview would be a more appropriate approach to understand their motivations. This presents opportunity for further experimental work.

6.5. Under-reporting Bias

In the previous section, we documented that there are huge cultural influences affecting the behavior of Chinese and American customers when they post online movie reviews. In this section, we study whether such behavioral differences affect the accuracy of the average of reviews as an indicator of the broader group's perception of quality. Given that prior research (Hu et al. 2006) has found the existence of under-reporting bias in United States consumers, causing online reviews to be a biased estimator of books' perceived quality, we study whether this is also true for movies and whether this varies across different cultures. We found that American reviews did indeed exhibit considerable under-reporting bias even for the bottom-ranked movies, but as expected we found much less under-reporting bias among Chinese reviewers.

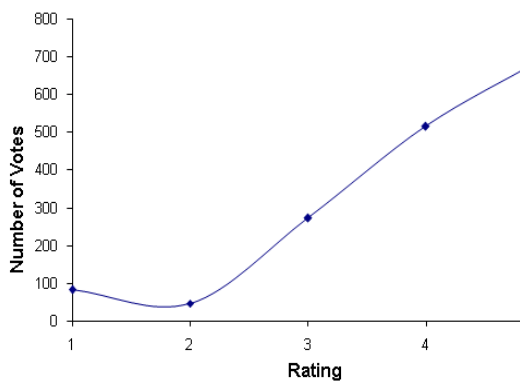


Figure 6.13: American Survey for Top-Ranked Movies

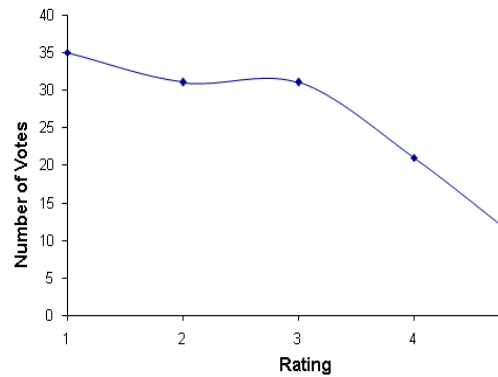


Figure 6.14: American Survey for Bottom-Ranked Movies

To obtain the set of reviews that are less likely to be subjected to under-reporting bias, we conducted an online survey approach. Survey requests were posted on Facebook and 247 survey responses were gathered.

In the survey, respondents were asked to report their ratings for several movies that they have viewed. Figure 6.13 and 6.14 show the survey results for Americans when assessing the top and bottom-ranked movies respectively. The survey results revealed patterns that are rather different to the results gathered from the online website, IMDB.com. Although the survey results for the top-ranked movies are somewhat similar to the online pattern, the survey results (Figure 6.14) obtained for the bottom-ranked movies are very different. If we remove the extremes, the pattern in Figure 6.14 follows an almost bell-shaped distribution for the bottom-ranked movies, in contrast to the pattern revealed in online in IMDB (Figure 6.11). By comparing the movie ratings in IMDB for the bottom-ranked movies with the results from the survey, we can conclude that the rating distributions are very different across these two channels for bottom-ranked movies. This confirms the existence of under-reporting bias in American movie reviewers, and such reporting bias does indeed cause online reviews to be a biased estimator of a product's true quality as perceived by the broader population of American movie-goers.

Next, we sought to verify whether such under-reporting bias exists in the other population we studied, that of the Chinese reviewers in Douban. Likewise, the existence of under-reporting bias among Chinese reviewers was examined using a set of controlled experiments in which all respondents were asked to report their ratings for several movies that they have viewed. Figure 6.15 and 6.16 show the survey results for the Chinese students when assessing the top and bottom-ranked movies respectively. The survey results revealed patterns similar to the results gathered from the online website, Douban.com. In particular, the results obtained from the online

movie reviews in Douban (Figure 6.6) and the results from our survey (Figure 6.16) follow a unimodal, almost normal distribution for the bottom-ranked movies. By comparing the movie ratings in Douban for the top-ranked movies (Figure 6.4) with the results from the survey (Figure 6.15), we can likewise conclude that the rating distributions are similar across these two channels for bottom-ranked movies as well. Furthermore, the mean difference between online ratings and offline ratings is insignificant (t-value=0.45 and p-value=0.6569).³³ Our interpretation for this is that there is far less under-reporting bias for Chinese online raters, and the overall online Chinese consumer opinion is a well representation of a product’s true quality, as measured by the average perception of the broader population of Chinese movie-goers.

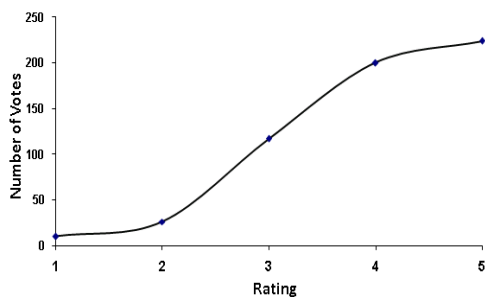


Figure 6.15: Chinese Students Survey for Top-Ranked Movies

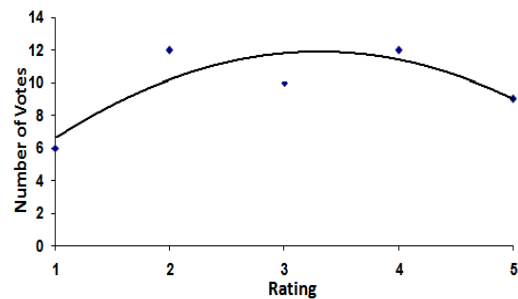


Figure 6.16: Chinese Students Survey for Bottom-Ranked Movies

We have shown that the extent of under-reporting bias does indeed vary across cultures. In comparison to the results obtained from the United States reviewers, the existence of under-reporting bias is less severe among Chinese movie reviewers. Hence, the average posted online ratings from Douban appear to be similar to those of the “silent consumers” who did not provide their ratings. However, in the case of

³³ The p-value for the F-test of equal variances is 0.7915, therefore we cannot reject the null-hypothesis that the underlying variances of the observations are equal.

IMDB, there seems to be far more online postings from movie-goers with the most extreme and indeed the most extremely negative views, and thus the set of posted reviews may not be truly and accurately representative of the “silent consumers” in the United States.

Most of the findings in this study are in accordance with expectations based on Hofstede’s cultural dimensions and his characterization of both American and Chinese cultures. As hypothesized, the differences in attitudes towards a movie — that is, differences in underlying assessments of the movie — had very different effects on the behaviors for online reviewing among collectivist and individualist populations. An explanation could lie in the influence of individualism in which individualists perceive that they are relatively free to follow their own wishes and outwardly express them. The fact that the percentages of reviews for the top-ranked and bottom-ranked movies in the U.S. sample are much higher than those of the Chinese sample is consistent with this assertion.

In terms of societal norms, collectivists display a much stronger adherence to the consensus of their communities, including of course the consensus of their online network communities. For the Chinese reviewers, there is concern for reconciliation, harmony, and balance. This may result in vague expression of personal emotions such as likes and dislikes.

To test if the results based on Chinese reviewers hold for other collectivist societies, we replicated our study in Singapore, which has been found to be more collectivist than individualist due to the Confucian heritage of the majority of the population (Hofstede 2001).

The Singapore Data

In this section, we examine how cultural elements influence the attitudes and intentions in the hybrid culture of Singapore. Although three-fourths of the Singapore population is Chinese, Singaporeans undergo a British system of education, with English being the main medium. Due to their colonial history, Singaporeans have been influenced by Western culture, but Eastern culture and values are also strong. Obedience, harmony and concern for reconciliation are important. Cultural factors have been shown to be mediators of attitude and behavior in Singapore (Tan and Farley 1987).

While Singaporean students are generally more exposed to Western values than their parents were, they still do possess traditional Chinese values as well. This perhaps explains why the results for the top-ranked movies (Figure 6.17) in the survey of Singaporean students were similar to those in Douban (Figure 6.4). In particular for the bottom-ranked movies in Figure 6.18, the experimental results revealed unimodal distribution with mostly moderate reviews, which is similar to the result for the Bottom 100 Movies in Douban.com, as in Figure 6.6.

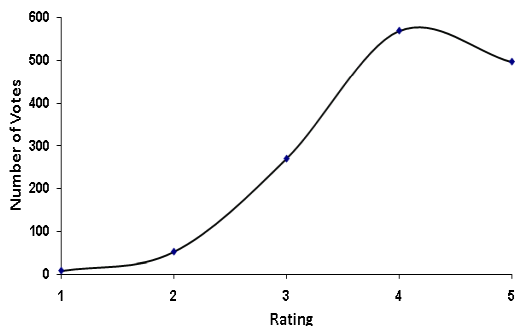


Figure 6.17: Singaporean Students Survey for Top-ranked Movies

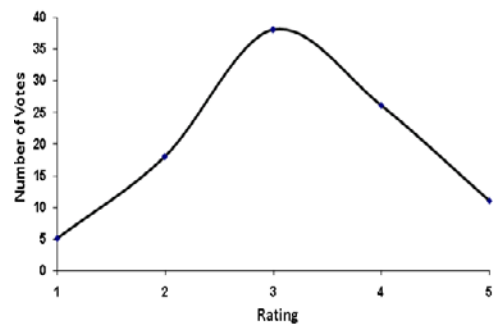


Figure 6.18: Singaporean Students Survey for Bottom-ranked Movies

80.5% of our survey participants have been to movie review websites, but, only 14.2% have posted ratings. When asked when they would be most likely to provide movie review ratings, 66.5% responded that they would if they liked the movie very much and only 29.1% responded that they would if they were very disappointed. This suggests that Singapore reviewers still exhibit characteristics of their parents' collectivist culture, similar to the behaviors we observed in our Chinese dataset. Once again we observe that reviewers in a collectivist culture are less likely to express their dissatisfaction. Similarly, by examining the data, we conclude that Singaporean reviewers are more reserved about giving the highest ratings even for the top-ranked movies. Once again, we observe support for Hypothesis 1.

6.6. Robustness Check

6.6.1. On a different dataset

In the previous sections, the analyses are based on the online dataset that were obtained through two rounds of data collection activities. For the first data collection round, reviews were collected on 1000 movies randomly selected from IMDB and 1000 randomly selected from Douban. We used these datasets to investigate the rating differences across these two websites. However, our conclusions might have been influenced by the differences between the movies selected from these websites for inclusion in our two datasets. To control the movie title effect, we therefore collected another round of data. For this round of data collection, we first gathered the movie reviews of the top-ranked 100 and bottom-ranked 100 movies in IMDB, then based on these movie titles we collected the related movie reviews from Douban.com. However, one might again argue that the top-ranked and bottom-ranked movies in

IMDB may not be top-ranked and bottom-ranked movies in Douban, which might once again have contributed to the observed differences.

Hence, we conduct a robustness check to ascertain if the results are consistent if we were to take the top and bottom-ranked movies in Douban and collect the corresponding reviews in IMDB. Since Douban does not have a bottom-ranked movie list, we had to scan manually through all the movies in Douban to find the bottom-ranked 100 movies, those with the lowest average rating. Then, based on the top-ranked and bottom-ranked movie list in Douban, we extract the corresponding movie ratings in IMDB and plot the rating distribution as shown in Figures 6.19 to 6.22. We find that the results were consistent with what we had previously obtained, and that once again extreme ratings are more prevalent among the online reviews written by U.S. movie raters, regardless whether the movies are top-ranked or bottom-ranked at either website.

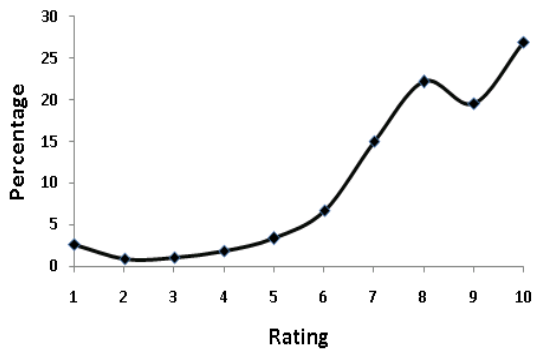


Figure 6.19: Corresponding (Top) IMDB Ratings

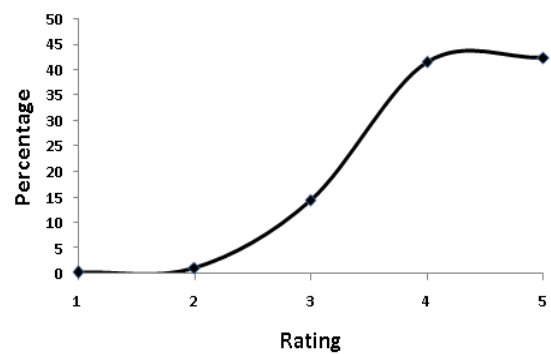


Figure 6.20: Douban Top-Ranked Ratings

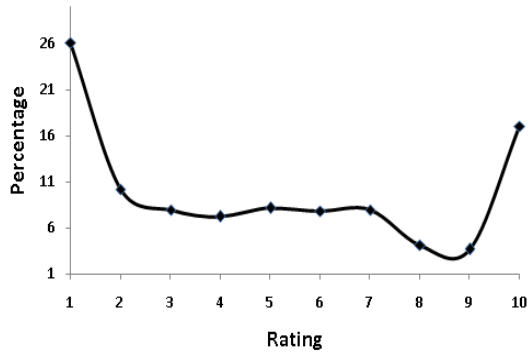


Figure 6.21: Corresponding (Bottom) IMDB Ratings

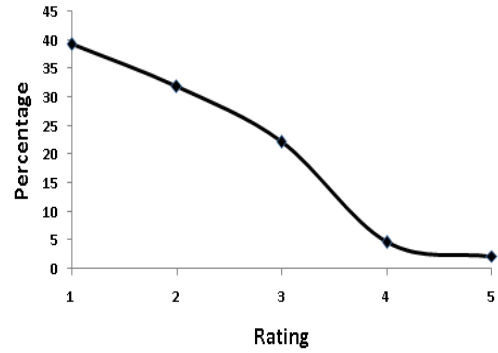


Figure 6.22: Douban Bottom-Ranked Ratings

6.7. Implications

This study contributes to our understanding of the role of social norms on individuals' behavior when writing online movie reviews and, we believe, to our understanding of the role of social norms in social networks more generally. Based on the empirical study conducted over three different population groups, from China, the United States, and Singapore, we find that under-reporting bias varies across cultures and cultural differences play a significant role in online reviewing behavior.

The main contributions come from applying Hofstede's (2001) cultural classifications, which predict certain differences in behaviors across cultures, and using these classifications to analyze differences in online behavior in a specific setting. We are not attempting to test Hofstede's theory, and we do not argue that all Chinese are collectivist, or that all Westerners are extreme individualists; we simply analyze millions of movie reviews posted online on websites that cater primarily to Chinese and American reviewers, and look for differences. The differences are indeed consistent with predictions based on Hofstede's theory: Western reviews are much more likely to be extreme, and their distribution tends to become more extreme over

time, while Chinese reviews tend to have a more bell-shaped distribution and newer additional posts are much more likely to be closer to the mean rather than more extreme. That is, in Western reviews we observe far more under-reporting among reviewers with average opinions. Such results are further validated by comparisons between the online and offline consumer reviews.

Movies have always been made principally for their home markets, but American movie producers in general hope for more global appeal and more global commercial success. This study indicates that online social reviewing behavior differs greatly from market to market, and might indeed lead a film's distributor to misjudge the size of a potential market abroad. In particular, a distributor based in one market will know how to interpret early reviews in his home market, but if he applies his home-market experience to interpreting the reviews from a foreign market he may be greatly misled. Reviewer behavior at home that indicates a moderately successful film might be associated with market failure, or with blockbuster success in another market. Thus, an American distributor might over-estimate the market in Singapore or in China, given the greater tolerance of reviewers, or, conversely, a Chinese distributor might under-estimate the market in America given the extreme behavior of some American reviewers. Most importantly, after comparing Figures 6.3 and 6.4, we realize that an American distributor might significantly *under-estimate* the market for a hit American movie in Singapore or China since Figure 6.4 does not exhibit the spike that correspond to American reviewers *over-reporting* of reviews of 9 and 10 for top-ranked movies. Likewise, a Chinese movie distributor might note a huge number of extremely negative reviews, an order of magnitude more than he might

expect in his home market, and conclude that the launch would be catastrophic. If online behavior is not representative of offline behavior, and if the differences between online and offline behavior vary by nation, then the information in online networks needs to be interpreted carefully before these reviews can be of use to either the community or marketers.

Finally, there are several limitations of this study. First, several parts of the study were performed using students; these portions of the study should be replicated with a non-student population. Second, we were unable to examine the individualists' motivation for posting reviews due to insufficient responses and we feel cultural differences may be at work here. Thus, we hope to gather sufficient responses in future to conduct further analysis on this part. Third, we did not examine the text comments for the posted reviews. It is likely that the extreme opinions of the Chinese are reflected only in the text comments instead of the numerical ratings. Therefore, future research could apply sentiment analysis techniques on the text comments to enable more comprehensive analyses. Also, it might be useful to see how similar results are in retail websites such as Amazon, eBay and others. Finally, further research on the behavior of Americans and of Chinese in online social networks and blogs not associated with commercial purposes would help to strengthen our understanding of cultural differences online.

This study investigates when the reported average of online ratings matches the perceived average assessment of the population as a whole, including the average assessments of both raters and non-raters. We apply behavioral theory to capture intentions in rating online movie reviews in two dissimilar countries – China and the

United States. We find that consumers' rating behaviors are affected by cultural influences and that they are influenced in predictable ways. Based on data collected from IMDB.com and DOUBAN.com, we found significant differences across raters from these two different cultures. Additionally, we examined how cultural elements influence rating behavior for a hybrid culture – Singapore. To study whether online consumer reviews are subjected to under-reporting bias, which is, consumers with extreme opinions are more likely to report their opinions than consumers with moderate reviews causing online reviews to be a biased estimator of a product's true quality, we compare the consumer reviews posted online with those from an experimental study. Our results shows that under-reporting is more prevalent among U.S. online network, thus online reviews are a better movie perceived quality proxy in China and Singapore than in the U.S.

Chapter 7.

Discussion and Conclusion

Online consumer product review is an emerging market phenomenon that is playing an increasingly important role in consumers' purchase decisions. This dissertation examines the value of online reviews by estimating its impact on product sales for firms (as discussed in Chapter 4 and 5) as well as informative worth for users of online reviews (in Chapter 6).

In essence, the studies conducted have empirically verified the importance and value of online reviews. Careful consideration on the design and structure of online reviews were noted so as to devise systematic and comprehensive study on the detailed effects of online reviews.

In addition to the differential impact of ratings and sentiments on sales, we also showed that there is a differential effect of sentiments expressed in the title of the review and the content of the review. The impact of content sentiments is substantially

larger than title sentiments. It is likely that in this case, customers are using title sentiments as a screening device but still would validate their choice by digging into the content sentiments. One possible implication for reviewers here is that they may need to pay attention to the way the title of the review is written. It should be crisp, and clearly pointing to the sentiments expressed in the full text so that it makes it attractive for potential buyers to look deeper at their review.

In dealing with the issue of manipulation of online reviews, this work has presented a simple but efficient technique to detect such activity. The empirical analysis suggests that the manipulation of online reviews to date is still able to influence sales positively. However, we have also found an abundant number of cases of manipulated reviews using our technique which is a cause for concern. If manipulative activity continues to expand, this may be detrimental to the credibility of online reviews in the long run.

Also for users of online reviews, it is not just the issue of online manipulation which they have to be aware of, but also the presence of under-reporting bias which may misrepresent the true perceived value of the product. Hence, both firms and individuals of online reviews will have to interpret the reviews carefully before making any decisions.

Using a large scale online review data and subsequent robustness check using different datasets for each study, the rigor of the research here present compelling and directional results for managers, marketers and online review users. The rest of this chapter sets out an agenda on some of the future work and extensions.

7.1. Future Work and Extensions

A couple of findings in this dissertation have offer insights that warrant our attention as they offer new avenues for further research. Based on the in-depth interviews with online shoppers and our experiential survey results, we show the relevance of rating and sentiments over different stages of the consumer decision making process. Future research, should not only look at the main effects of ratings and sentiments of online reviews, but also consider their *mediating* effects.

We made a preliminary attempt at exploring this through an observational study (five students), where we presented a search scenario (find travel guides to New Zealand) and observed on how they went through the search process on Amazon.com. Once the search results are returned, the participant skims through a large set of relevant books and identify a subset to evaluate in greater depth. From our interview with the participants, they normally select books with high ratings to narrow their choice and click on further to read the reviews. Finally, the customer makes his/her final choice and proceeds with the purchase of the desired book(s). We further conducted a survey with 156 (100 male, 56 female) Amazon.com users. These were all undergraduate students from a major business school in the United States. Participation was voluntary and no monetary incentive was given. Respondents are given a scenario of buying a travel guide on New Zealand from Amazon.com. Suppose the keywords typed in are “travel guide on New Zealand” they are presented with the screen shot of the search results that would appear on Amazon.com. A series of questions were asked as to the relative importance of numerical ratings and text reviews during the search, evaluation and purchase. 58% of the respondents felt that

numerical ratings are important in the early stages of search and awareness whereas 65% of the respondents felt that text sentiments are important when making a purchase (Figure 7.1a). This is just a preliminary look at this issue and is only indicative of the possible role of ratings and sentiments in the purchase process. Future work using experiments, eye-tracking or other related methods (Chandon et al 2006, Wedel and Pieters 2008) will help in getting a better understanding of this process that is suggested here (Figure 7.1b). The results in this study suggests the relevance of ratings and sentiments may be different over the course of search, evaluation and purchase and we hope that these findings would be an impetus for future research on this timely and important topic.

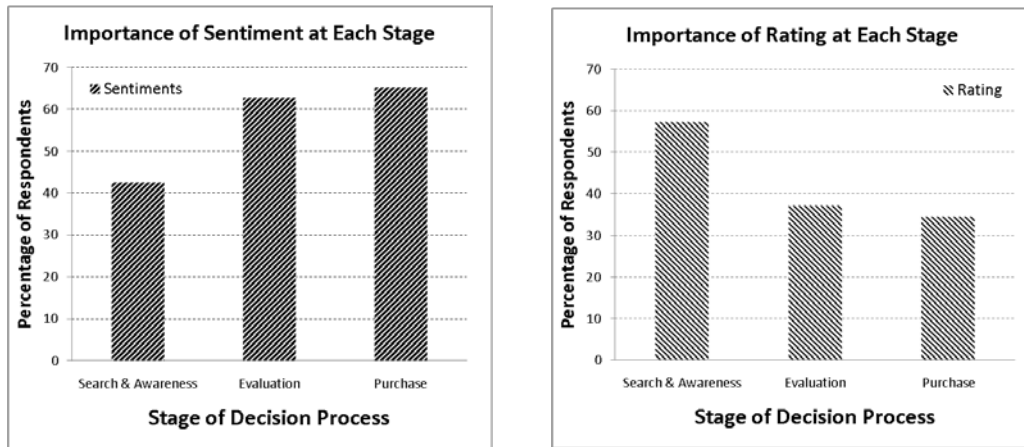


Figure 7.1a: Importance of Numerical Ratings and Text Sentiments in Purchase Process

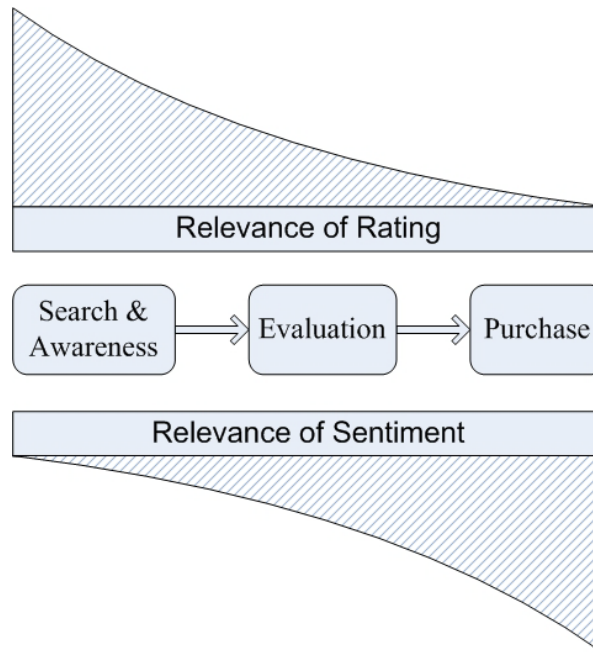


Figure 7.1b: Ratings and Sentiments and their Relevance in Stages of the Decision Making Process

Also, given the limited attention each consumer is able to spend, it is unlikely that a consumer will process all the available reviews before purchase (Forman et al. 2008). Consumers have to use heuristics to select a subset of the reviews to read rather than processing all the reviews systematically (Forman et al. 2008). Thus, the use of eye tracker may help to enhance our understanding on what types of reviews consumers usually focus on. In addition, using the eye-tracking focal points, we can estimate the importance of the kind of sentiment words customers pay attention to.

The work of Forman et al (2008) found significant effect of reviewer disclosure on sales. As we were unable to collect this information in our study, we could not include this variable in our model. Ghose and Ipeiritis (2010) found that average helpfulness of the review was impacted by reviewer disclosure. We have incorporated

average helpfulness in our model and hope that some of the disclosure impact may be captured by this variable. However, future work should incorporate this important variable. Another limitation is that the current study does not explicitly control for the heterogeneity in scale usage (Rossi et al 2001). As respondents may vary in their usage of scale, these differences may impart biases on regression estimates. Future work should identify the existence of heterogeneity in scale usage and correct for it using procedures suggested by Rossi et al. (2001).

In many cases of the WOM literature, researchers also tend to make causal arguments about the relationship between reviews and sales (Chevalier and Mayzlin 2006). All these studies implicitly assume that online consumer ratings reflect consumer's evaluation of the product. However, one of the major difficulties related to estimating the impact of reviews on sales rank is the endogeneity problem – books with higher intrinsic quality tend to have better reviews, so it is hard to determine whether the positive review or the high quality of a book is responsible for its high sales rank. In the literature, most papers generally do not consider the endogeneity issue, which means the findings of a positive influence of reviews on sales rank may be spurious.

More recent work has addressed this problem using several approaches. Chevalier and Mayzlin (2005) use “difference-in-differences” methods to eliminate fixed effects over time and across different websites (Amazon and Barnes & Noble). Elberse and Eliashberg (2003) examine the causal relationship between movie advertising and revenue using panel data analysis to eliminate the fixed effects of movie quality. Zhang and Dellarocas (2006) build a structural model for the word-of-

mouth diffusion process and exploit the weekly changes in revenue to control for the unobservable intrinsic quality and other time-invariant factors of movies. Given the concern on endogeneity, one extension in our work is to investigate whether the influence of reviews can be identified using a multiple equation model. Preliminary results can be found in Appendix Section A.1.

Using multiple equation models (see Appendix A.1), our preliminary study provides a clearer understanding on the inter-relationship between ratings, sentiments and sales rank. We find that there is a differential role of rating and sentiments on sales rank. The ratings effect on sales rank is mostly indirect through sentiments while sentiments effect on sales rank is mostly direct. This is consistent with our bootstrap mediation analysis, thus indicating the role of sentiments as a mediator. Sales rank on the other hand does not have impact on ratings and sentiments contemporaneously. We also find that ratings have a contemporaneous effect on sentiments but sentiments do not affect ratings. The mediating role of sentiments suggest a possible sequential decision making process where ratings play an important role in early stages during search and text sentiments play an important role in evaluation and choice.

We again showed that there is a differential effect of sentiments expressed in the title of the review and the content of the review. The impact of content sentiments is substantially larger than title sentiments. Another interesting finding in our study is that moderate, ordinary sentiments have a stronger impact than strong sentiments. Ordinary negative and ordinary positive sentiments have a greater impact on sales than strong positive and strong negative sentiments. This interesting result can be

explained in the context of online reviews. It appears customers seem to find more value in sentiments in more moderately worded reviews and they seem to find the ordinary negative and ordinary positive sentiments to have the greatest value.

Interestingly, we find that the accessibility to online reviews (Hsee 1996, Shah and Oppenheimer 2008) does create a differential impact on sales. Since the most helpful and most recent reviews are easily accessible to customers, their impact on sales is much larger than the average impact of all reviews. It seems that customers do rely on the most recent and most helpful reviews to make their evaluation and choice. This is an area that requires further research. Researchers attempting to extract sentiments from text reviews may examine whether a subset of reviews that consumers use may just be sufficient to see the impact on sales. Further research on understanding the consumer behavior in the rich user-generated environment in terms of search strategies, heuristics used in evaluating and making choices is going to be important for firms to design their websites to make it easier and more helpful. As our current empirical work is based on aggregate data, future work should explore this with individual level data.

Given the importance of online reviews, another future work would be to examine the dynamics of online review. Compared to traditional consumer surveys that estimate consumer evaluation of a certain product, online consumer ratings have many very distinct features: firstly, online reviews are usually retained since the first review; secondly, unlike offline surveys, each respondent answers the question independently, in an online setting, review posters see all the existing reviews. Therefore, it is likely that reviewers may be influenced by the information presented

on the webpage. Thus, in presence of strong network effects, it is crucial to understand the review patterns and sentiments of customers so as to devise effective business strategies. Potential questions to be addressed in future would be: 1) what are the underlying patterns of online reviews over time? 2) How does it differ for products of different popularity? and 3) How does it differ for products of different categories?

In this regard, if subsequent reviews for products are merely restating the early reviews, the usefulness and impact of the subsequent reviews would be marginal as compared to the early reviews. However, when subsequent reviews do have different attributes from the early reviews, firms have to treat reviews differently and construct different strategies depending on the life cycle of the product. This future study will present a guide on reviewers' attitude and can assist companies to obtain a holistic view of how online reviews evolve over time so as to better predict sales.

Last but not least, what are some of the important properties of the next generation sentiment mining system needed to address the specific types of questions that management researchers will be examining in future? As for any scientific problem, there is a need to first formalize the problem. As sentiment analysis is still a difficult task and to effectively study the sentiments in detail is extremely challenging especially for large scale datasets, the next generation of sentiment mining system would need the following properties:

- 1) ability to automatically extract sentiments for terabytes of datasets; and

- 2) ability to score sentiments on a multi-point scale at a document level, sentence level or for different features, depending on the research question.

Figure 7.1c presents a sketch of the next generation sentiment analysis system. Usually, sentiment mining requires the following steps (though technical details may differ). The first step is to construct a dictionary. The second step is to extract words that express positive or negative sentiments for a product (or product feature). Finally a summary of the sentiments is produced.

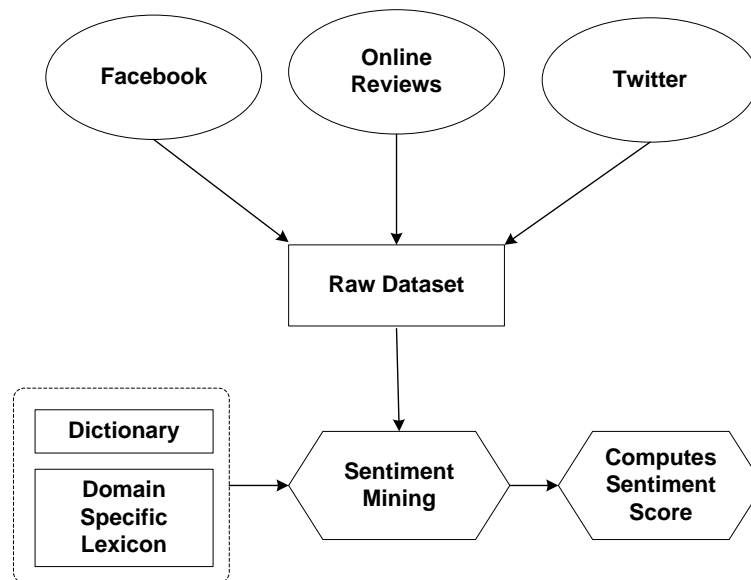


Figure 7.1c: A sketch of the next generation sentiment analysis system

A foreseeable task based on the current trend is the ability to perform sentiment mining for large-scale datasets either from online reviews, Twitter, Facebook and so on. However, unlike online reviews, the textual comments in Facebook and Twitter are may be shorter, more abbreviated and less formal. Due to the peculiarities of the language used in social media site like Facebook and Twitter, there is a need to include a domain specific dictionary. This will help to identify certain commonly used terms in social media (particularly those used by the new

generation of population also known as “Generation Y”) to increase the accuracy of results. For example, abbreviations or sentiment terms like “*very farnie* (funny)”, “LOL (laugh out loud)”, “*eeeewwwww* (expressing disgust)” and so on have to be added in a domain specific dictionary.

In conclusion, the next generation system design of sentiment analysis requires the capability to handle large-scale dataset and have a domain specific dictionary to help increase the accuracy of the results. This will serve as a common framework to unify different research directions and identify what the main tasks of sentiment analysis are, their inputs and outputs and how the resulting outputs may be used in practice by managers.

REFERENCES

Alexa Internet. *Top Sites China*.

([://www.alexa.com/site/ds/top_sites?cc=CN&ts_mode=country&lang=none](http://www.alexa.com/site/ds/top_sites?cc=CN&ts_mode=country&lang=none)) 2008.

Ajzen, I. 1985. From intentions to actions: A theory of planned behavior, in Kuhl, J. and J. Beckmann (eds.), *Action Control: From Cognition to Behavior*.

Ajzen, I. 1988. *Attitudes, Personality, and Behavior*, The Dorsey Press, Chicago, Illinois.

Ajzen, I. 1991. The theory of planned behavior, *Organizational Behavior and Human Decision Processes*, 50, 179-211.

Anderson, E. W. 1998. Customer satisfaction and word of mouth, *Journal of Service Research*, 1 (1) 5-17.

Anderson, E.W. 1998. Customer satisfaction and word of mouth. *Journal of Service*. 1(1) 5-17.

Archak, N., A. Ghose, P.G. Ipeirotis. 2007. Show me the money! Deriving the pricing power of product features by mining consumer reviews. *Proceedings 13th International Conference Knowledge Discovery and Data Mining*, 56-65.

Basuroy, S., S. Chatterjee, S. A. Ravid. 2003. How critical are critical reviews? The box office effects of film critics, star-power, and budgets. *Journal of Marketing*. 67(4) 105-117.

Beneish, M.D. 1999. The detection of earnings manipulation, *Financial Analysts Journal*. 55 (5) 24-36.

- Belsley, D., E. Kuh, R. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Collinearity*. Wiley, New York.
- Bickart, B., R. M. Schindler. 2001. Internet forums as influential sources of consumer information. *Journal of Interactive Marketing*. 15(3) 31-40.
- Bikhchandani S., D. Hirshleifer, I. Welch. 1992. A theory of fads, fashion, custom and cultural change as information cascades. *Journal of Political Economy*. 100(5) 992-1026.
- Bond, R. and P. B. Smith. 1996. Cross-cultural social and organizational psychology, *Annual Review of Psychology*, 47, 205-235.
- Brynjolfsson, E., Y. Hu, M. Smith. 2003. Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety. *Management Science* 49(11) 1580-1596.
- Chandon, P., J. W. Hutchinson, E. Bradlow, S. H. Scott. 2006. Measuring the value of point-of-purchase marketing with commercial eye-tracking data. *INSEAD Business School Research Paper* No. 2007/22/MKT/ACGRD. Available at SSRN: <http://ssrn.com/abstract=1032162>
- Chatterjee, P. 2001. Online Reviews: Do consumers use them? *Advances in Consumer Research* 28 (1) 129-133.
- Chen, C. 2002. Robust Regression and Outlier Detection with the ROBUSTREG Procedure, in the *Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

- Chen, P., S. Wu. 2004. The impact of online recommendations and consumer feedback on sales. *Proceedings of 24th International Conference on Information Systems*, 711-724
- Chen, Y., J. Xie. 2004. Online consumer reviews: A new element of marketing communications mix. *Management Science*, 54(3) 477-491.
- Clemons, E.K., G. Gao, L. Hitt. 2006. When online review meets hyperdifferentiation: A study of craft beer industry. *Proceedings of 39th Annual Hawaii International Conference on System Sciences*.
- Chevalier, J. A., A. Goolsbee. 2003. Measuring prices and price competition online: Amazon.com and BarnesandNobel.com. *Quantitative Marketing and Economics*. 1(2) 203-222.
- Chevalier, J. A., D. Mayzlin. 2006. The effect of word of mouth online: Online book reviews. *Journal of Marketing Research*. 43(3) 345-354.
- Chintagunta, P.K., S. Gopinath, and S. Venkataraman. 2010. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science* Published online before print May 27, 2010.
- Clemons, E. K. 2008. How information changes consumer behavior and how consumer behavior determines corporate strategy. *Journal Management Information Systems*. 25 (2) 13-40.
- Clemons, E.K. and G. Gao. 2008. Consumer informedness and diverse consumer purchasing behaviors: Traditional mass-market, trading down, and trading out into the long tail, *Electronic Commerce Research and Applications*. 7 (1) 3-17.

- Clemons, E.K., S. Barnett, A. Appadurai. 2007. The future of advertising and the value of social network websites: Some preliminary examinations. In C. Dellarocas and F. Dignum (eds.), *Proceedings of the 9th International Conference on Electronic Commerce*, Minneapolis, 267-276.
- Das, S. R., M. Chen. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*. 53(9) 1375-1388.
- Dave, K., S. Lawrence, D. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of 13th International Conference World Wide Web*.
- David, S., T.J. Pinch 2005, Six degrees of reputation: The use and abuse of online review and recommendation systems, working paper retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=857505.
- Dellarocas, C. 2003. The digitization of word-of-mouth: Promise and challenges of online feedback mechanisms. *Management Science*. 49(10) 1407–1424.
- Dellarocas, C., N. Awad, X. Zhang. 2004. Exploring the value of online reviews to organizations: Implications for revenue forecasting and planning. *Proceedings 25th International Conference on Information Systems*, New York: ACM press, 379-386.
- Dellarocas, C., R. Narayan. 2006. A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth. *Statistical Science*. 21(2) 277-285.
- Duan, W., B. Gu, A. Whinston. 2008. Do online reviews matter? - An empirical investigation of panel data. *Decision Support Systems*. 45(4) 1007-1016.

- Ehrens, S., A. Markus. 2000. Amazon.com: There's an "R" in e-tailing. Epoch Partners Consumer Internet Company Report (November 13).
- Elberse, A., J. Eliashberg. 2003. Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures. *Marketing Science*. 22(3) 329–354.
- Ess, C., and F. Sudweeks. 2005. Culture and computer-mediated communication: Toward new understandings, *Journal of Computer-Mediated Communication* 11 (1) ([.indiana.edu/vol11/issue1/ess.html](http://indiana.edu/vol11/issue1/ess.html)).
- Fishbein, M. and I. Ajzen. 1975. *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*, Addison-Wesley, Reading, MA.
- Forman, C., A. Ghose, B. Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*. 19(3) 291-313
- Garsten, B. 2005. *Saving Persuasion: A Defense of Rhetoric and Judgment*, Harvard University Press, Boston.
- Ghose, A., P.G. Ipeirotis. 2010. Estimating the socio-economic impact of product reviews: Mining text and reviewer characteristics. Forthcoming in *IEEE Transactions on Knowledge and Data Engineering*.
- Ghose, A., P.G. Ipeirotis, A. Sundararajan. 2006., Evaluating pricing strategy using ecommerce data: evidence and estimation challenges, *Statistical Science*. 21 (2) 131-142.
- Ghose, A., P.G. Ipeirotis, A. Sundararajan. 2007. Opinion mining using econometrics: A case study on reputation systems. *Proceedings 44th Annual Meeting of the Association for Computational Linguistics*.

- Ghose, A., M. Smith, R. Telang. 2006. Internet Exchanges for Used Books: An Empirical Analysis of Product Cannibalization and Welfare Impact, *Information Systems Research*. 17(1), 3-19,
- Godes, D., D. Mayzlin. 2004. Using online conversation to study word of mouth communication. *Marketing Science*. 23(4) 545-560.
- Godes D. and J. Silva. 2006. The Dynamics of Online Opinion, working paper.
- Gruhl, D., R. Guha, R. Kumar, J. Novak, A. Tomkins. 2005. The predictive power of online chatter. *Proceedings 11th International Conference Knowledge Discovery in Data Mining*, New York, NY, USA 78–87.
- Guernsey, L. 2000. Suddenly, everybody's an expert on everything, *The New York Times*, 3rd February.
- Gujarati, D. N. 2003. *Basic Econometrics*. 4th ed. New York: McGraw–Hill, Inc.
- Gurun, U.W., A.W. Butler 2009. Don't believe the hype: Local media slant, local advertising and firm value, retrieved from: [://papers.ssrn.com/sol3/papers.cfm?abstract_id=1333765](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1333765).
- Haubl, G., V. Trifts. 2000. Consumer decision making in online shopping environments: the effects of interactive decision aids. *Marketing Science*. 19(1) 4-21.
- Harmon A. 2004. Amazon glitch unmask war of reviewers, *The New York Times*, 14th February.
- Hatzivassiloglou, V., K. R. Mckeown. 1997. Predicting the Semantic Orientation of Adjectives. *Proceedings 35th Annual Meeting of the Association for Computational Linguistics*, 174-181.

- Herr, P.M., F. R. Kardes, J. Kim. 1991. Effects of word of mouth and product attribute information on persuasion: an accessibility-diagnostics perspective. *Journal of Consumer Research*. 17(4) 454–462.
- Holmes, D.I. 1994. Authorship attribution, *Computers and the Humanities*. 28 (2) 87-106.
- Hofstede, G. 1980. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations*. Sage Publications, Thousand Oaks CA.
- Hofstede, G. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations*, 2nd Ed. Sage Publications, Thousand Oaks CA.
- Hofstede, G. and M. H. Bond. 1988. The Confucius connection: From cultural roots to economic growth, *Organizational Dynamics*. 16 (4) 4-21.
- Hsee, C. K. .1996. The Evaluability Hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*. 67 (3), 247–57.
- Hu, M., B. Liu. 2004. Mining and Summarizing Customer Reviews. *Proceedings of the Tenth ACM International. Conf. Knowledge Discovery and Data Mining*, 168-177.
- Hu, N., I. Bose and L. Liu. 2010a. Manipulation in digital word-of-mouth: A reality check for book reviews, forthcoming in *Decision Support Systems*.
- Hu, N., L. Liu and V. Sambamurthy. 2010b. Fraud Detection in Online Consumer Reviews, forthcoming in *Decision Support Systems*.

- Hu, N., Pavlou, P.A, J. Zhang. 2006. Can Online Reviews Reveal a Product's True Quality? Empirical Findings and Analytical Modeling of Online Word-of-Mouth Communication. *Proceedings of the Seventh ACM Conference on Electronic Commerce*, Ann Arbor, Michigan, 324-330.
- Huang, J. H., Y. F. Chen. 2006. Herding in Online Product Choice. *Psychology and Marketing*. 23(5) 413-428.
- Ito, T.A., J. T. Larsen, N. K. Smith, J. T. Cacioppo. 1998. Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology*. 75 (4) 887-900.
- Kahn, K.F., P.J. Kenney. 2002. The slant of the news, *American Political Science Review*. 96 (2) 381-394.
- Klare, G.R. 2000. The measurement of readability: useful information for communicators, *ACM Journal of Computer Documentation*, 24 (3) 107-121.
- Kim, J. Y., K. Ryu. 2003. Yes-Men and No-Men: Does defiance signal talent? *Journal of Institutional and Theoretical Economics*. 159(3) 468-490.
- Lanier, J. 2010. *You Are Not a Gadget: A Manifesto*. Alfred A. Knopf, New York.
- Lewitt, S., C. Syverson. 2005. Market distortions when agents are better informed: The value of information in real estate transactions, working paper.
- Li, C., and Bernoff J. 2008. *Groundswell: Winning in a World Transformed by Social Technologies*. Harvard Business Press, Boston.
- Li, X., L. Hitt. 2008. Self-Selection and information role of online product reviews. *Information Systems and Economics*. 19(4) 456-474.

- Li, X. and Hitt, L. 2010. Price Effects in Online Product Reviews: An Analytical Model and Empirical Analysis, forthcoming in *MIS Quarterly*.
- Liu, Y. 2006. "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of Marketing*, 70(3) 74-89.
- Liu, B., M. Hu, J. Cheng. 2005. Opinion Observer: Analyzing and comparing opinions on the web. *Proceedings of 14th International Conference of World Wide Web*, 342-351.
- Mahajan, V., E. Muller, R. Kerin. 1984. Introduction strategy for new products with positive and negative word-of-mouth. *Management Science*. 30(12) 1389 - 1404.
- Majumdar, S., D. Kulkarni, C. Ravishankar. 2007. Addressing Click Fraud in Content Delivery Systems. *Infocom. IEEE* . [://www.cs.ucr.edu/~smajumdar/infocom07.pdf](http://www.cs.ucr.edu/~smajumdar/infocom07.pdf).
- Madden, T.J., Ellen P.S., and I. Ajzen. 1992. A comparison of the theory of planned behavior and the theory of reasoned action, *Personality and Social Psychology Bulletin*, 18 (1) 3-9.
- Maslow A.H. 1943. A theory of human motivation, *Psychological Review*, 50, 4, 370-96.
- Mayzlin, D. 2006. Promotional chat on the Internet, *Marketing Science*. 25 (2) 155-163.
- Metwally, A., D. Agrawal, and A. E. Abbadi. 2005. Using association rules for fraud detection in web advertising networks. *In Proceedings of the 31st international Conference on Very Large Data Bases*, 169-180.
- Moe, W., M. Trusov. 2010. Measuring the Value of Social Dynamics in Online Product Ratings Forums, working paper.

- Morinaga, S., K. Yamanishi, K. Tateishi, T. Fukushima. 2002. Mining product reputations on the web. *Proceedings of 8th International Conference Knowledge Discovery and Data Mining*, 341-349.
- Nakata, C., K. Sivakumar. 2001. Instituting the marketing concept in a multinational setting: The role of national culture, *Journal of the Academy of Marketing Science*, 29 (3) 255-275.
- Nelson, P. 1970. Information and consumer behavior. *Journal of Political Economy*. 78(2) 311 - 329.
- Paasche-Orlow, M.K., H.A. Taylor, F.L. Brancati. 2003. Readability standards for informed-consent forms as compared with actual readability, *New England Journal of Medicine*. 348 (8) 721-726.
- Pang B., L. Lee. 2004. A Sentimental Education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, 271–278.
- Pang B., L. Lee. 2005. Seeing Stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics*, 115–124.
- Pang B., L. Lee, S. Vathiyathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing*.
- Pavlou, P.A. 2002. What drives electronic commerce? A theory of planned behavior perspective, *Best Paper Proceedings of the Academy of Management Conference*, Denver, CO, August.

- Pavlou, P.A., C. Lin. 2002. What drives electronic commerce across cultures? A cross-cultural empirical investigation of the theory of planned behavior, *Journal of Electronic Commerce Research*, 3 (4) 240-253.
- Poulton, E.C. 1989. *Bias in quantifying judgments*. Hillsdale, NJ:Erlbaum.
- Preacher, K., A. Hayes. 2008. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*. 40 (3), 879-91.
- Reddy S. K., V. Swaminathan, C. M. Motley. 1998. Exploring the Determinants of Broadway Show Success. *Journal of Marketing Research*. 35(3) 370-383.
- Rossi, P.E., Z. Gilula, G.M. Allenby. 2001. Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association*. 96(453) 20-31.
- Roychowdhury, S. 2004. Manipulation of earnings through the management of real activities that affect cash flow from operations, Unpublished dissertation, University of Rochester.
- Schnapp, M., T. Allwine. 2001. Mining of Book Data from Amazon.com. *UCB/SIMS Web Mining Conference*
at [://www.sims.berkeley.edu:8000/resources/affiliates/workshops/webmining/Slides/ORA.ppt](http://www.sims.berkeley.edu:8000/resources/affiliates/workshops/webmining/Slides/ORA.ppt)
- Senter, R.J., E.A. Smith. 1967. Automated readability index, retrieved from: <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=AD0667273>.
- Senecal, S., J. Nantel. 2004. The influence of online product recommendations on consumers' online choices, *Journal of Retailing*. 80 (1) 159–169.

- Shah, A. K., D. M. Oppenheimer. 2009. The path to least resistance: Using easy-to-access information. *Current Directions in Psychological Science*, 18, 232-236.
- Siau, K., J. Erickson, F. F. Nah. 2010. Effects of National Culture on Types of Knowledge Sharing in Virtual Communities, forthcoming in *IEEE Transactions on Professional Communication*.
- Sondergaard, M. 1994. Research note: Hofstede's consequences: A study of reviews, citations and replications. *Organization Studies*. 15(3) 447-456.
- Stone, P. J., D. C. Dunphy, M. S. Smith, D. M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*, MIT Press, Cambridge, MA.
- Tan, C.T., and J. U. Farley. 1987. The impact of cultural patterns on cognition and intention in Singapore. *Journal of Consumer Research*. 13 (4) 540-544.
- Taylor, S. and P. A. Todd. 1995. Understanding information technology usage: A test of competing models, *Information Systems Research*. 6 (3) 144-176.
- Tourangeau, R., L. J. Rips, K. Raskinski. 2000. *The Psychology of Survey Response*. Cambridge University Press.
- Tsang, A.S.L., G. Prendergast 2009. Is a "star" worth a thousand words?: The interplay between product-review texts and rating valences. *European Journal of Marketing*. 43(11) 1269 - 1280
- Turney, P. D. 2002. Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of 40th Annual Meeting on Association for Computational Linguistic*, 417-424.

- Turney, P. D., M. L. Littman. 2002. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*. 21 (4) 315-346.
- Wedel, M., R. Pieters. 2008. Eye Tracking for Visual Marketing. *Foundations and Trends in Marketing*. 1 (4), 231–320.
- White, E. 1999. Chatting a singer up the pop charts, *The Wall Street Journal*, 5th October.
- Wu, F. and Huberman B. 2007. Novelty and collective attention, in *Proceedings of the National Academy Science*. 104 (45) 17599-17601.
- Yaffee, R. A. 2002. *Robust Regression Analysis: Some Popular Statistical Package Options*, (<http://www.nyu.edu/its/statistics/Docs/RobustReg2.pdf>)
- Zakaluk, B.L., S.J. Samuels. 1988. *Readability: Its Past, Present, and Future*, International Reading Association, Newark.
- Zhang, X., C. Dellarocas. 2006. The Lord of the Ratings: How a Movie’s Fate is Influenced by Reviews?,” in *Proceedings of the 27th International Conference on Information Systems (ICIS)*, 1959-1978.
- Zhang, X., Dellarocas, C., N. F. Awad. 2004. “Estimating Word-of-Mouth for Movies: The Impact of Online Movie Reviews on Box Office Performance,” *Workshop on Information Systems and Economics (WISE)*, College Park, MD.
- Zhao, X., J. Lynch, Q. Chen. 2010. Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis. *Journal of Consumer Research*. 37, 197-206.

Appendix

A.1 Multiple equation analysis

This section presents a preliminary multiple equation analyses on a panel dataset to analyze the relationship between rating, sentiments and sales rank. We collected a panel data set of books sold on Amazon.com from September 2005 to January 2006 using its Web Service (AWS). We initially chose 10,000 books randomly to gather sales, and review information. Of these 10,000 books we found that 4,405 books had text reviews that we can capture in September 2005 and January 2006. For each item, we collected the title, when the book was released, Amazon's retail price and sales rank (which we used as a proxy for sales). In addition, we collected information on reviews, such as total number of reviews (volume), the numerical rating from which average (valence) and variance can be computed, helpfulness of review, and the original text of reviews from which sentiments can be extracted and scored.³⁴ We also computed sentiment scores for the title of the review as well as the content of the review. In consideration that consumers do not read all reviews, we collected ratings and sentiment information on the most recent reviews and most helpful reviews³⁵.

The summary statistics are provided in Table A.1a.

³⁴ For some items, the reviews that were posted too long ago were no longer available and could not be collected.

³⁵ Amazon.com provides this type of information on the first page of each item with the most helpful reviews provided on the left/center of the page and most recent reviews on the right hand side of the page. We thank one of the reviewers for suggesting that we look at this in addition to the average ratings and sentiments of all reviews. We have looked at the top five and ten most helpful reviews and the five and ten most recent reviews for each item when they are available. The results between the five and ten most helpful and recent reviews did not differ much. We have thus chosen to present the results for five most helpful and recent reviews.

Table A.1a: Summary Statistics of Panel Dataset from Amazon.com

Variable	September 2005		January 2006	
	Median	Mean (SD)	Median	Mean (SD)
Price	16.29	18.15 (22.73)	15.72	19.06 (16.60)
Sales Rank	3955.00	36,653.31 (98,796.62)	4317.00	53,196.36 (135,516.17)
Age (Days)	149.00	421.24 (803.69)	297.00	657.02 (945.03)
Number of Reviews	12.00	42.68 (161.36)	12.00	43.85 (168.75)
Average helpful ratio	0.40	0.41 (0.16)	0.17	0.20 (0.22)
Average Rating	4.16	3.88 (0.78)	4.00	3.72 (1.21)
Average sentiment score	3.66	3.71 (0.42)	3.78	3.73 (0.72)
Variance of rating	1.29	1.62 (1.13)	1.41	1.70 (1.62)
Variance of sentiments	0.69	0.70 (0.36)	0.58	0.69 (0.65)
Average Title Sentiment	3.61	3.62 (0.60)	3.67	3.70 (1.08)
Average Content Sentiment	3.70	3.80 (0.40)	3.84	3.76 (0.62)
Most Helpful Rating	4.00	3.88 (1.05)	4.00	3.61 (1.39)
Most Helpful Sentiment Score	4.04	3.99 (0.48)	3.63	3.68 (0.54)
Recent Rating	4.00	3.85 (0.85)	3.80	3.66 (0.91)
Recent Sentiment Score	3.98	3.77 (0.51)	3.64	3.66 (0.53)
Sample Size		4405 books		4405 books

The following three equation model captures not only affects of various product and user generated characteristics on sales, but also the interrelationships between ratings, sentiments and sales.

Salesrank :

$$\ln(SR)_{jt} = \alpha_0 + \alpha_1(AR)_{jt} + \alpha_2(AS)_{jt} + \alpha_3 \ln(P)_{jt} + \alpha_4 \ln(Age)_{jt} + \alpha_5 \ln(TR)_{jt} + \alpha_6(AH)_{jt} + \alpha_7(VR)_{jt} + \alpha_8(VS)_{jt} + \mu_j + \varepsilon_{SRjt}$$

Sentiment :

$$AS_{jt} = \beta_0 + \beta_1 \ln(SR)_{jt} + \beta_2 \ln(SR)_{j(t-1)} + \beta_3(AR)_{jt} + \beta_4 \ln(P)_{jt} + \beta_5 \ln(Age)_{jt} + \beta_6 \ln(TR)_{jt} + \beta_7(AS)_{j(t-1)} + v_j + \varepsilon_{ASjt}$$

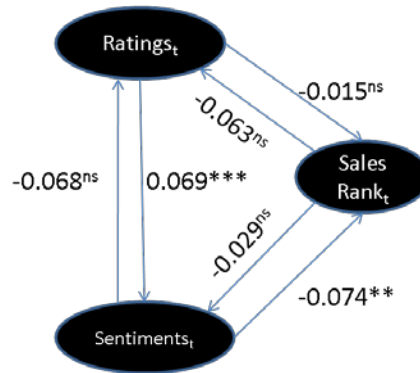
Rating :

$$AR_{jt} = \gamma_0 + \gamma_1 \ln(SR)_{jt} + \gamma_2 \ln(SR)_{j(t-1)} + \gamma_3(AS)_{jt} + \gamma_4 \ln(P)_{jt} + \gamma_5 \ln(Age)_{jt} + \gamma_6 \ln(TR)_{jt} + \gamma_7(AR)_{j(t-1)} + \omega_j + \varepsilon_{ARjt}$$

where,

- j = 1, ..., N book items.
- P_{jt} = price of book j at time t
- Age_{jt} = age of book j at time t
- TR_{jt} = total number of reviews of book j at time t
- AH_{jt} = average helpfulness ratio of book j at time t
- SR_{jt} = sales rank of book j at time t
- AR_{jt} = average rating of book item j at time t .
- AS_{jt} = average sentiment of the book item j at time t .
- VR_{jt} = variance of rating for book item j at time t .
- VS_{jt} = variance of sentiment for book item j at time t .
- $SR_{j(t-1)}$ = sales rank for book item j at time $t-1$.
- $AR_{j(t-1)}$ = cumulative average rating of book item j by time $t-1$.
- $AS_{j(t-1)}$ = cumulative average sentiment of the book item j by time $t-1$.

μ_j , v_j and ω_j are the product-level fixed effects for the three equations respectively to control for unobserved heterogeneity across products and ε_{SRjt} , ε_{ASjt} , ε_{ARjt} are the residual error terms.



Significance: *** $p < .001$; ** $p < .01$; * $p < .05$

Figure A.1a: The Results of the Inter-relationship between Ratings, Sentiments and Sales Rank

The three equation non-recursive model is estimated using three-stage least squares with fixed effects³⁶. The model estimates (standardized) appear in Table A.1b³⁷.

³⁶ The fixed effects procedure is used to eliminate the biasing influence of unobserved fixed book-specific effects. Three stage least squares estimation method takes care of potential correlation of error terms across the three

Figure A.1a above highlights the interrelationship between sales rank, sentiments and ratings. Examining the relationships between the three endogenous variables, we find that the impact of average ratings on sales is not significant whereas the impact of average sentiments on sales is negative and significant (-0.074). Our results indicate that the impact of ratings on sales rank is mostly indirect through sentiments and the impact of sentiments on sales rank is mostly direct. Much of previous research which looked at just numerical ratings found a direct impact on sales (Chevalier and Mayzlin 2006, Forman et al. 2008)³⁷. Ghose and Ipeiritis (2010) who looked at electronic products in their research and had some measures of sentiments (namely writing style) found that the ratings had a significant effect on sales in only one of the three product categories (in the presence of average subjectivity of reviews). Our finding of the impact of numerical ratings on sales being mostly indirect and through sentiments is an interesting and an important one. This finding suggests a potential sequential nature of consumer decision making. Due to the nature of the complex task of searching and purchasing in an online environment, consumers may use different strategies to lessen the burden of their cognitive effort. The way that they may do this is by using ratings as a way to screen potential items and use text reviews to evaluate the limited set of screened items to make the final choice. Although this is an interesting finding, caution should be exercised as this is based on

equations. Hausman-Wu endogeneity test indicated both Price and Total Number of Reviews were independent of the contemporaneous error terms.

³⁷ Models without fixed effects showed an adjusted R^2 of 0.50 for sales rank equation, 0.56 for sentiments equation and 0.58 for ratings equation. It appears that accounting for quality differences in books in our model has helped the explanatory power.

³⁸ These are the two pieces of research which looked at Amazon.com books in their analysis and are most relevant for comparison. Other researchers investigating different product categories found mixed impact of ratings on sales. Whereas Liu (2006), Duan et al. (2008) found no significant impact of ratings on box-office of movies, Dellarocas et al. (2004), Chintagunta et al. (2010) found positive impact. Clemons et al. (2006) looking at beer and Moe and Trusov (2010) looking at bath fragrances and beauty products found significant effect ratings on sales.

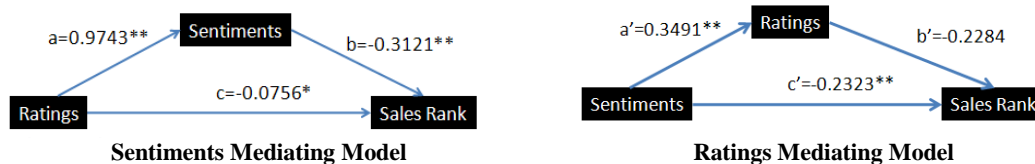
aggregate data and not individual level data. Future work should explore this further to identify the mechanism and the role played by ratings and sentiments on the online decision making process³⁹.

Table A.1b: Model Estimates (3SLS Fixed Effects Model)

	SR_{jt}	AS_{jt}	AR_{jt}
$\ln(SR)_{jt}$ (Sales Rank)	-	-0.0298	0.0626
$\ln(SR)_{j(t-1)}$ (Lag Sales Rank)	-	-0.0402*	-0.0006
AR_{jt} (Average Rating)	-0.0148	0.0689***	-
AS_{jt} (Average Sentiment)	-0.0742**	-	-0.0681
$\ln(P)_{jt}$ (Price)	0.3310***	-0.0579***	-0.0221
$\ln(Age)_{jt}$ (Age)	0.1995***	-0.0102	-0.0857***
$\ln(TR)_{jt}$ (Total Review)	-0.8305***	-0.0442***	-0.1044***
AH_{jt} (Average Helpful)	-0.1068***	-	-
VR_{jt} (Variance of Rating)	0.0063	-	-
VS_{jt} (Variance of Sentiment)	0.0013	-	-
$AR_{j(t-1)}$ (Average Rating at t-1)	-	-	0.2839***
$AS_{j(t-1)}$ (Average Sentiment at t-1)	-	0.1815***	-
Intercept	0.0230***	0.0367***	0.0156
Adjusted R ²	0.98	0.93	0.87
N	4405	4405	4405

Standardized coefficients. Significance: *** $p < .001$; ** $p < .01$; * $p < .05$

³⁹ To examine these interrelations between ratings, sentiments and sales rank further, we performed mediating analysis suggested by Zhao, Lynch and Chen (2010) to look at whether either of these two consumer review measures plays a mediating role. What is the role that ratings and sentiments play in impacting sales? To understand if ratings or sentiments play a mediating role, we estimate two models: one with sentiments as the mediating variable and the second with ratings as the mediating variable. It involves estimating the coefficients, a, a', b, b', c and c'. c and c' indicate the direct impact and the product a x b and a' x b' will provide the indirect impact of ratings and sentiments on sales rank respectively. The significance of the indirect effect will determine the mediating role that these two measure of consumer reviews play. Zhao et al. 2010 recommend using Preacher and Hayes (2008) bootstrapping procedure to determine the significance of the indirect effects.



The estimates of the models are presented in the figure above. The indirect effects of ratings ($a \times b = (0.9743) \times (-0.3121) = -0.3041$) was significant (the 95% bootstrap confidence interval was -0.3668 and -0.2431). The indirect effects of sentiments ($a' \times b' = (0.3491 \times -0.2284) = -0.0797$) was not significant (the 95% bootstrap confidence interval was -0.1150 and 0.0430). The results show that 80% of the total effect of ratings on sales rank is indirect, whereas the total effect of sentiments on sales rank is direct. This suggests the mediating role played by sentiments. Ratings thus, influence sales more through sentiments than directly.

Alternative models.

As mentioned earlier, consumers may sample some reviews of items and not go through all the reviews. As Amazon.com presents the most helpful reviews and the most recent reviews on the first page of an item, consumers may choose to use them when evaluating and making a purchase decision. It would be interesting to see what the impact on sales rank will be if we use valence information (ratings and sentiments) from only the most helpful and the most recent reviews.⁴⁰

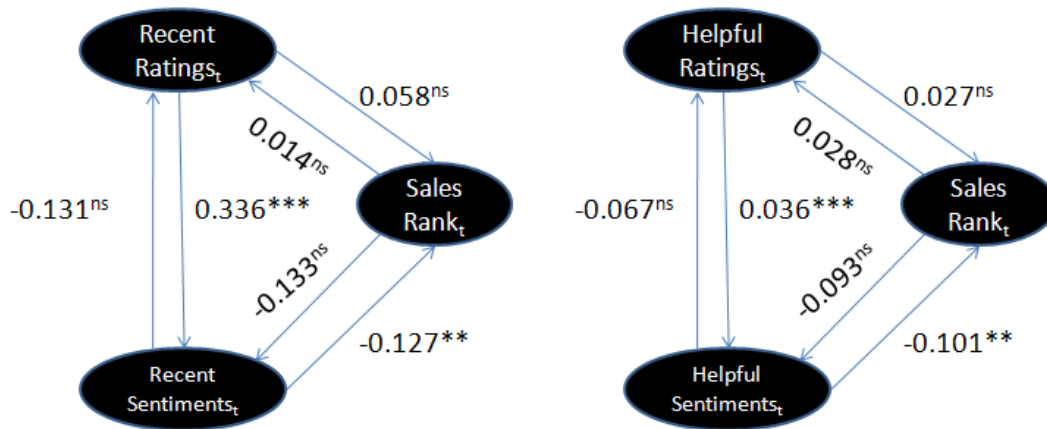
Many times, the title of the review presents a summary view of what is in the full text of the review. Customers look at the titles of the reviews to get a feel for what the review might say and then decide to take the decision to read the text of the review. To see if there is a differential impact of the sentiments in the title and the content of the review, we decomposed the total sentiment score into that based on just the title of the online review and the one based on the body (or content) of the review. To understand how the sentiments expressed in the title and the body of the review affect sales rank, we estimated a model where we had average sentiments from the title of the review and the body of the review included (instead of the overall average sentiments).

In the computation of the summary sentiment score for each review we gather information on the number of strong positive, strong negative, ordinary positive and ordinary negative sentiments expressed. It will be interesting to see their differential impact on sales. So we estimate models with various combinations of these sentiments.

⁴⁰ The averages are taken for the five most helpful reviews and the five most recent reviews for each book item. We also estimated models using 6-10 most helpful / recent reviews. We obtain similar results. We discuss the results for five most helpful and five most recent reviews.

Impact of Most Recent and Most Helpful Reviews.

Figure A.1b presents the estimates of the interrelations between most helpful/most recent ratings, sentiments and sales rank. We have substituted most helpful/most recent ratings/sentiments for average ratings/sentiments in the original model⁴¹. The use of the most recent or most helpful reviews does not change the qualitative interpretation of the effects. The ratings (whether most recent or most helpful) still seem to have an indirect impact on sales rank and sentiments (whether most recent or most helpful) still seem to have a significant direct impact on sales rank. However, some of the standardized coefficient estimates are larger in these models. For example, the direct effect of most helpful sentiments (-0.101; $p < 0.01$) and most recent sentiments (-0.127; $p < 0.01$) on sales rank is larger than the corresponding impact of average sentiments (-0.074; $p < 0.01$).



Significance: *** $p < .001$; ** $p < .01$; * $p < .05$;

Figure A.1b: The Results of the Interrelationship between Most Helpful/Most Recent Ratings, Sentiments and Sales Rank

⁴¹ The overall results of these models parallel the results reported earlier for the average ratings/sentiments model. The estimates for the full model for all these alternate models can be obtained from the authors

This is an interesting finding as it indicates that the most evident and accessible set of reviews on Amazon.com (namely the most helpful and most recent) play a significant role in determining sales. The mental effort required by consumers in reading through a large number of reviews is minimized by sampling the most recent and most helpful reviews to make their evaluation and choice. This is an area that requires further research. Researchers attempting to extract sentiments from text reviews may examine whether a subset of reviews that consumers use may just be sufficient to see the impact on sales. Also, from the perspective of Amazon.com or other similar sites, the way and the kind of user-generated information that is presented to the consumers may make a difference. In this case, Amazon.com provides these two types of reviews in an easily viewable way for consumers. If they present a sample of the most positive and most negative reviews in an easy accessible manner, will they show a strong impact on sales as well?

From the heuristic perspective, the information that has more accessibility has dominant impact on judgment and decision making (Hsee 1996, Shah and Oppenheimer 2009) and our results on the most helpful and most recent reviews seems to suggest this. Understanding the consumer behavior in the rich user-generated environment in terms of search strategies, heuristics used in evaluating and making choices is going to be important for firms to design their websites to make it easier and helpful.

Impact of Title and Content Sentiments.

We examine the impact of ratings and sentiments taking into account the structure of the review. Both the title and content sentiments have significant negative impact on sales rank. We find that the sentiments in the content (body) of the review have a larger impact than the sentiments in the title of the review (-0.2304; $p < .001$; vs. -0.0129; $p < .01$). Title may convey some information that is useful, but the customers seem to be paying more attention to the sentiments expressed in the content of the review. The impact of ratings on sales rank is relatively smaller (-0.005; $p < .05$) as compared to the sentiments in the title and content. Consistent with our prior results, sales rank does not affect sentiments and rating. The relationship between sentiments and rating again is one-way i.e. ratings affect sentiments but not vice versa.

Impact of Strong and Ordinary Sentiments

Next, we examine how sentiments of different strengths may affect sales rank. Examples of strong positive sentiments are words like “excellent” and “awesome” while strong negative sentiments are words like “terrible” and “awful”. Examples of ordinary positive sentiments are words like “nice”, “satisfactory” and ordinary negative sentiments are words like “redundant”, “dislike”⁴². We estimate four different models, using the same structure as in the original model, but with different

⁴² The strong positive/negative score or ordinary positive/negative score for the i^{th} review is calculated using the following formula:

$$\frac{\textit{senti_part}_i}{SP_i + SN_i + OP_i + ON_i},$$

where $\textit{senti_part}_i \in \{SP_i, SN_i, OP_i, ON_i\}$

The strong positive (SP) / negative (SN) score or ordinary positive (OP) / negative (ON) score for each product is obtained from the content of each review. The final strong positive/negative score or ordinary positive/negative scores of a product is the average of all the strong positive/negative score or ordinary positive/negative scores over all the reviews received by that product item respectively.

pairs of sentiments. We examine the impact of positive sentiments (strong positive and ordinary positive), negative sentiments (strong negative and ordinary negative), strong sentiments (strong positive and strong negative) and ordinary (ordinary positive and ordinary negative) sentiments. As four sets of sentiments are not linearly independent, we can use only at most three of the sentiments in the model. We chose to use combinations of the sentiments to get a better understanding of the impact on sales rank.

Table A.1c: Compiled Results from Models Using Different Sentiments and their Impact on Sales Rank

	Model 1	Model 2	Model 3	Model 4
<i>AR_{jt}</i> (Average Rating)	-0.0230*	-0.006*	-0.005	-0.008*
<i>ASN_{jt}</i> (Strong Negative)	0.0144**			
<i>AOP_{jt}</i> (Ordinary Positive)	-0.0224***			
<i>AON_{jt}</i> (Ordinary Negative)	0.1580***			
<i>ASP_{jt}</i> (Strong Positive)		-0.0660***	-	-
<i>AOP_{jt}</i> (Ordinary Positive)		-0.1227***	-	-
<i>ASN_{jt}</i> (Strong Negative)		-	0.0751***	-
<i>AON_{jt}</i> (Ordinary Negative)		-	0.1028***	-
<i>ASP_{jt}</i> (Strong Positive)		-	-	-0.0021*
<i>ASN_{jt}</i> (Strong Negative)		-	-	0.0571***

Standardized coefficients. Significance: *** $p < .001$; ** $p < .01$; * $p < .05$

Thus, in the original model, we replace the AS_{jt} variable in the sales rank equation first with three of the sentiments namely, average strong negative sentiments (ASN_{jt}), average ordinary positive sentiments (AOP_{jt}) and ordinary negative sentiments (AON_{jt}). We estimated models with the following sets of sentiments: 1) ASN_{jt} , AOP_{jt} , AON_{jt} 2) ASP_{jt} , AOP_{jt} 3) ASN_{jt} , AON_{jt} 4) ASP_{jt} , ASN_{jt} . The compiled results of the

coefficients of strong and ordinary sentiments are presented in Table A.1c⁴³. As the results on other variables do not change qualitatively, we have compiled the key results of the sentiment variables from our models in Table A.1c.

A consistent picture that gets observed here is that ordinary sentiments seem to have a stronger impact on sales rank than strong sentiments (Model 1, 2 and 3). The impact of ordinary positive sentiments is stronger than strong positive sentiment (Model 2). Similarly, ordinary negative sentiments have a stronger impact than strong negative sentiments (Model 3). This is a very interesting result as it appears that contrary to what one would expect that strong positive or negative sentiments are not as impactful as ordinary positive and negative sentiments. The impact on sales rank is greater for strongly negative sentiments compared to strongly positive sentiments Model 4. Chevalier and Mayzlin (2006) found that the one-star reviews had relatively larger coefficients than five-star reviews. Their interpretation is that relatively rare one-star reviews carry a lot of weight with consumers. Our findings extend this further by showing that the relatively balanced sentiments (not using either strong positive or negative terms) are generally more valued.

A.2 Dataset

In this dissertation, the datasets used for empirical analysis are available from <http://www.mysmu.edu/phdis2006/noisian.koh.2006/> or may be requested from hunan@smu.edu.sg.

⁴³ The results of the other variables in the model are qualitatively similar to the ones reported earlier in Table A.1b and are not presented here for brevity.